

A Survey of Information Retrieval Vendors

Robert J. Kuhns

Sun Microsystems Laboratories
2550 Garcia Avenue
Mountain View, CA 94043

Introduction

This report is a survey of vendors that develop and market information retrieval (IR) technology. The objective of this survey is to provide information for those who want an overview of text retrieval and document management companies, their products, and their indexing and searching capabilities. More specifically, it presents summary information for each corporation surveyed regarding its business, markets, products, and technology. Although the conclusion contains a discussion of trends, commonalities, and differences among the vendors in the information retrieval industry, this report does not make direct comparisons or competitive assessments between individual vendors.

Since the report contains current summary information on vendors and their product offerings, it can be used as a resource for decisions involving software acquisitions and technology assessments.

All information contained in this report is publicly available.

Scope of the Survey

This survey covers 23 vendors that sell or license text retrieval technology. (Inso Corporation is an exception in that it licenses linguistic-based software to Original Equipment Manufacturers (OEMs) that can be used in information retrieval applications.) The actual choice of vendors

was made based on their presence (size of customer base and/or corporate longevity) in the IR community or by their novel approach to text processing. Vendors were not excluded based on their target platforms for their software.

It should be noted that this survey is not exhaustive. A few vendors are not included because they did not provide information on their products after several attempts in trying to get that information. Moreover, in order to bound the survey, academic and industrial research groups are not included in the survey. For the same reason, companies offering search facilities for the World Wide Web exclusively are not discussed. Companies developing text categorization and data extraction software were also omitted because their orientation and technology are different from IR methods.

The vendors surveyed all have their corporate headquarters either in the U.S. or Canada. Due to lack of success in obtaining information, there are no Asian or European vendors represented in this survey.

The companies that are covered in this report are: askSAM Systems, CLARITECH Corporation, ConQuest Software, Inc., Cuadra Associates, Inc., Data Retrieval Corporation, Dataflight Software, Inc., Dataware Retrieval Corporation, Excalibur Technologies Corporation, Fulcrum Technologies, Inc., HNC Software, Inc., Information Access Systems, Inc., Information Dimensions, Inc., Inso Corporation, Odyssey Development, Inc., Open Text Corporation, Personal Library Software, Inc., Quality Information Systems, Inc., re:Search International, TextWare Corporation, Thunderstone Software, TMS, Inc., Verity, Inc., and Virginia Systems Software Services, Inc.

Form of Survey

The body of the survey consists of a set of corporate profiles, one for each vendor. The general structure of each profile is as follows.

Title

(Is the full name of the company.)

Headquarters

(Includes mailing address, telephone and Fax numbers, email address, and URL for company's World Wide Web page.)

Corporate Fact Sheet

(Includes sections on the business, the markets, and possibly sample applications and key customers of the vendor.)

Products

(Includes sections on the vendor's core technology (search and retrieval engine) and related products based on the core technology. The supported platforms are also listed.)

Technical Fact Sheet

(Includes sections on indexing methods, searching functionality, and any other pertinent information related the vendors technology.)

Differentiating Features

(Includes a short description of some of the key aspects of a vendor's technology. This section is optional.)

Miscellaneous

(Includes information about the company that does not fit into any of the other sections. This section is optional.)

While the overall structure of the vendors are similar to one another, the amount of detail and internal structure differs from one vendor to the other. This is especially true for the section entitled "Technical Fact Sheet" which could have been written in a rigidly formatted style using a checklist for indexing and search features. However, this approach would have missed many of the subtleties that exist from one vendor's technology to the next. Also, different vendors use the same term to refer to similar, but slightly different, functionality (especially, searching features) which would be difficult to capture in a template style structure.

There are two main reasons for differences in the level of detail among the profiles. As noted above, the information in the survey is publicly available, and consequently, vendors vary on how much information they are willing to share without executing a non-disclosure agreement. Furthermore, the complexity of the technology itself impacts the amount of detail that exists.

askSAM Systems

Headquarters

askSAM Systems
PO Box 1428
Perry, FL 32347

Tel: 800.800.1997 or 904.584.6590
Fax: 904.584.7481
email: info@asksam.com
WWW: <http://www.asksam.com/asksam.htm>

Corporate Fact Sheet

Business

askSam develops and markets information retrieval software.

Markets

askSam does not direct its marketing at any particular industry or group of users. However, about a third of askSam users are in the legal community. Journalists, writers, and publishers also comprise a large segment of askSam's user base. Currently, askSam is looking at opportunities on the Internet.

Sample Applications

askSam was used for searching the testimony in the Iran Contra hearings.

Products

Core Technology

askSam is an off-the shelf search and retrieval product.

Related Products

Read-only Runtimes allows users to distribute documents with search capabilities.

askSam Network Version provides askSAM's capabilities over a local area network (LAN).

askSam's Electronic Publisher allows users to distribute information electronically. It includes search capabilities, a report writer, and an unlimited viewer license.

ReadIris OCR Module scans text into databases. It includes askSam's search capabilities.

Quotes on Line DOS contains more than 12,000 quotations from more than 4,000 sources.

Platforms

askSam runs under Microsoft DOS and Windows. It requires 4MB of RAM and 7MB of free disk space.

Technical Fact Sheet

Indexing

In its latest release, askSam supports full-text indexing. Additionally, users can access structured information by adding fields to the searchable database. Information that is imported can be automatically placed in the appropriate fields according to pre-defined fields or templates. For instance, if email is imported into askSam, the fields TO:, FROM:, and DATE: which are part of the email format can be used for restricting searching in an askSam database to specified fields.

With OLE (Object Linking and Embedding), askSam permits non-text objects to be included in the askSam database.

AskSam offers several pre-defined templates including address, calendar, clippings, email, notes, questionnaire, phone directory, and to-do lists.

Searching

Boolean, wildcard, case sensitive or insensitive, date, numeric, hypertext, and proximity searches are supported. askSam does not include any stemming utility.

CLARITECH Corporation

Headquarters

CLARITECH Corporation
319 South Craig Street, Suite 200
Pittsburgh, PA 15213-3726

Tel: 412.621.0570
Fax: 412.621.0569

email: info@clarit.com
WWW: <http://www.clarit.com>

Corporate Fact Sheet

Business

CLARITECH develops and markets indexing and retrieval software.

CLARITECH was founded in 1992 to commercialize information retrieval software under development since 1988 at the Laboratory for Computational Linguistics at Carnegie Mellon University.

Markets

CLARITECH markets its software to OEMs, commercial information management software vendors, system integrators, and customers engaged in document and information management, such as developing historical archives.

CLARITECH can also be used for categorization and tagging applications such as news dissemination and indexing for on-line services.

Sample Applications

National Technology Transfer Center (NTTC)—NTTC is a clearinghouse for information on advanced technology projects in the United States. For NTTC, CLARITECH has developed a system to scan incoming technical documents, index them, issue English queries, and retrieve information as compound text/image documents. The system will be distributed to federal research laboratories.

Heinz Electronic Library Interactive Online System (HELIOS)—HELIOS is an electronic archive containing the professional papers of the late Senator H. John Heinz, III. In cooperation

with the Carnegie Mellon Libraries and the Laboratory for Computational Linguistics, CLARITECH is developing a system that will: manage over 1 million compound documents; include different interfaces for clerical workers, archivists, and end-users; and provide tools for researching, browsing, and retrieving documents in the archive. HELIOS will be integrated with the Libraries at Carnegie Mellon and will be available to authorized users via the internet.

The Netherlands Organization for Applied Scientific Research (TNO)—TNO, in conjunction with Digital Equipment Corporation, Cap Gemini Innovation, and Image Group Europe, is developing an information service for education and research materials in the field of environmental problems. The search engine for the project (called MOOI) is CLARIT. In preparation for MOOI, CLARIT was adapted to Dutch in three days and to French in four days.

CLARITECH has also participated in the Text Retrieval Conference (TREC), which is a project to evaluate the effectiveness of information retrieval technologies. It is co-sponsored by the national Institutes of Standards and Technology (NIST) and the Advanced Research Projects Agency (ARPA).

Products

Core Technology

CLARIT is a tool that indexes texts in terms of linguistic structures coupled with statistical routines. CLARIT accepts natural language queries and has the capability of augmenting these queries with words and phrases extracted from the document collection. CLARIT can be used for information filtering and routing applications, as well as for assigning subject and other tags to documents.

Related Technology

CLARIT Digital Archive System is a set of tools for creating and managing document collections including historical archives, corporate libraries, reference libraries, and specialized collections. It provides a client/server architecture which enables access to archives using a standard WWW browser (e.g., Mosaic or Netscape), or using a proprietary Windows-based interface. The Digital Archive System consists of the CLARIT Archive Builder and CLARIT Archive Publisher.

The Archive Builder uses scanning and Optical Character Recognition (OCR) to convert paper documents into page-image and electronic-text formats. It includes a user-interface for reviewing the page images and OCR text, and for adding archival categorization and annotations.

The Archive Publisher lets users search the full contents of the digital library and review page images. The Archive Publisher has two components, namely, the CLARIT Indexing and Retrieval Server and the CLARIT End-User Interface.

The Index and Retrieval Server prepares documents for retrieval and then services queries from

multiple simultaneous users using the CLARIT search engine.

The End-User Interface permits users to search the archive using English queries and archival organization (e.g., series and subseries), see a list of results sorted by archival categories, and view the page images of the retrieved documents.

The CLARIT Information Management System provides interactive use of the CLARIT engine, document display, and browsing functions.

The CLARIT Developer's Toolkit is an object-oriented application programming interface (API) that permits the integration of CLARIT modules into other applications. System function calls are available using C or C++.

Platforms

The CLARIT Indexing and Retrieval Server is available on DEC Alpha (OSF/1), DEC MIPS (Ultrix), IBM RS/6000 (AIX), Sun™ SPARC™ workstations (SunOS™ 4.1.3 or Solaris™ 4.2), and Windows NT.

The CLARIT Archive Builder consists of a scanner, high-resolution monitor, OCR software, and an IBM-compatible PC.

The CLARIT Archive Publisher requires one or more servers running UNIX® from DEC, Sun, or IBM, or a Windows NT workstation. Users access the system using a standard WWW browser, or via a proprietary CLARIT interface running on a PC under Windows NT.

Benchmarks

On a DEC Alpha 3000/400 (133Mhz and 128MB of RAM), CLARIT indexes at a rate of 80-100 MB per hour. Index size is about 50-85% of the original data.

On the same machine and for a 250MB database, a query of two sentences in length executes in 0.25 seconds, while a query involving a paragraph requires 0.50 seconds.

CLARIT can accept up to 8GB of source text to build a single index. A larger database can be indexed and searched across several server machines.

Technical Fact Sheet

Indexing

Indexing in CLARIT involves document preparation, morphological analysis, parsing, and scoring. CLARIT can handle most text formats with text preparation involving the conversion of text into CLARIT format by adding demarcation symbols for document boundaries, titles, and other fields.

Based on the CLARIT lexicon and any specialized lexicons, each word in the text is reduced to its root form. Noun phrases are then extracted and weights are assigned to the noun phrases in a document.

Noun phrase extraction involves head noun (or noun compound) and all pre-nominal modifiers. Post nominal constructions, e.g., prepositional phrases and relative clauses, are processed for noun phrases and are not attached to head nouns.

CLARIT treats every single-word and multi-word term as independent with respect to weighting. An inverse document frequency/ term frequency (IDF_xTF) term weighting routine is used to calculate weights for each independent term. Each document is then represented as a vector of weighted terms.

Searching

Queries are represented as vectors similar in construction to the vectors representing documents. Relevance rankings are produced by a vector modeling distance function between query and document vectors.

CLARIT accepts English queries. Boolean queries, wildcards, and proximity operators are supported as post natural language constraints.

Documents are ranked based on relevance.

Users have the option to modify weights for particular terms in a query.

Search terms can also be augmented (i.e., new terms suggested) automatically, based on the terms in the document collection.

Users can use relevant passages of previous hits for subsequent searches. The terms that match the new query, i.e., selected passage, can be highlighted for inspection.

Hyperlinks between related documents can be automatically generated.

Differentiating Features

A key feature of CLARIT is its use of natural language technology, namely, morphological analysis and noun phrase identification. In both indexing and query processing, words are decomposed into their root forms and noun phrases are determined. Unlike most systems that produce an inverted index of individual words and pre-specified phrases, CLARIT includes noun phrases that are automatically extracted from text in its indexes. CLARIT also generates a list of words and phrases which can be used to enhance or augment queries.

CLARIT can be adapted to other languages with a suitable lexicon and grammar.

ConQuest Software, Inc.

Headquarters

ConQuest Software, Inc.
10440 Little Patuxent Parkway, Suite 800
Columbia, Maryland 21044

Tel: 410.740.8800
Fax: 410.740.8810
email: Conquest@cq.com
WWW: Not available.

Corporate Fact Sheet

Business

ConQuest Software, Inc. develops and markets indexing and text retrieval technology for standalone, networked, and client/server environments. ConQuest claims to be the first text retrieval vendor to employ semantic networks as the basis of its search capability.

ConQuest was purchased by Excalibur Technologies in a stock deal that closed on July 20, 1995. It is expected that a combined offering of products from ConQuest and Excalibur will be available in mid-1996. Users will be able to search images, audio, video, drawings, or text using either the bit-level pattern recognition engine from Excalibur or semantic networks from ConQuest.

ConQuest was established in 1989.

Markets

ConQuest applications include competitive intelligence, computer-aided new drug applications, government intelligence, information categorization, message handling, litigation support, medical research, news indexing, filtering, and routing, and regulatory information management.

Sample Applications

Calspan Corporation—Calspan and ConQuest have been selected to provide a document content analysis and retrieval system to the National Air Information Center of the United States Air Force.

E-Systems/Garland Division—ConQuest supplied the engine for ECLIPS, E-Systems business

intelligence system.

The Analytical Sciences Corporation (TASC)—ConQuest is used as the key component in TASC's internal business intelligence system. ConQuest is also a partner in the Information Refinery, a subsidiary of TASC developing intelligence systems for government and commercial customers.

Other partners and customers of ConQuest include Blueridge Technologies, Infonautics' Homework Helper on Prodigy, Lotus Development Corporation, Motorola Incorporated, National Institutes of Health, Physicians' Online, the Veterans Administration Quality Management Bookshelf (on CD-ROM), the Federal government's Global Change-Assisted Search for Knowledge project, and U.S. intelligence agencies.

Products

Core Technology

ConQuest Text Management System is a system for indexing, searching, and retrieving documents. It consists of a query program to accept and execute queries, an index program to index new or updated documents, a library manager to manage the organization of text files, and a dictionary editor to maintain dictionaries and private concepts. The Text Management System also includes ConQuest's semantic network.

Related Products

The ConQuest product line consists of ConQuest Development and Integrator Systems, Third Party Products, and the ConQuest Profiling and Dissemination System.

ConQuest Development and Integrator Systems is a development environment including Application Programmer Interfaces (APIs) for integrating ConQuest's information and text management system with users' applications.

Third Party Products are turnkey applications for business intelligence, categorization, information dissemination, and information retrieval in the Lotus Notes environment.

The Profiling and Dissemination System provides for a real-time and archival information categorization and routing system. Users develop profiles using search terms or selected passages from documents, and evaluate them against a test database. Once developed, these profiles act as filtering agents for delivering relevant information to users.

The Profiling and Dissemination System uses full word indexing, a profile maintenance component that allows users to alter their profiles, as well as an API for integrating the system with other applications.

The profiling system operates differently from other search and retrieval systems. To minimize retrieval time, the profile database is stored in RAM. Each profile requires about 2000 bytes and, thus, an application with 100,000 profiles would use about 200MB of RAM. Each document is then processed against the index of profiles rather than matching the each profile against a document.

ConQuest has benchmarked the system on a Sun SPARCstation™ 20 with 256MB of RAM. With 1,000 profiles at an average profile length of 20 words, the system categorized over 165MB of incoming information per hour. At 90,000 profiles, the system categorizes 7.2MB per hour.

Conquest currently has a relational database management system (RDBMS) bridge and a Visual Basic interface to its technology in beta release.

Platforms

ConQuest products can run standalone, networked, or in a client/server environment including:

Servers

DEC Alpha OSF
HP UX
IBM RS-6000
Sequent Dynix/PTX
SGI IRIX
Sun Solaris 2
Sun OS 4.1.3
Other UNIX System 5 Version 4
Windows/NT

Clients

Windows
Macintosh
Motif

Networks

Novell, NFS™ (Sun's distributed computing file system), and UNIX TCP/IP

Technical Fact Sheet

Indexing Methods

Indexing consists of a set of language processing modules followed by a pair of components for preparing text for querying. Documents (ASCII text) are first decomposed into basic parts by a document parser. Each structured part of the document is tokenized, and the tokens are then

passed to a morphological analyzer which determines roots of inflected forms. The next module is actually a chunker, in that it takes the root forms and determines multi-word phrases, e.g., the word "digital" immediately preceding the word "library" would be analyzed as a single unit, namely, "digital library." This module also handles idioms as well.

Two other modules supporting indexing apply before an index is created. At this point, stop words are removed and the document is indexed. The first of these normalizes dates and numbers, while the second removes the stop words.

There are no practical limits to document collections to be indexed since document collections can be divided into libraries which can be indexed and searched simultaneously.

In a set of tests, Conquest has benchmarked their indexing component up to 90MB per hour.

Index size is typically 50% of the original collection.

The indexing representation is proprietary to ConQuest.

Searching Methods

Query Processing

As with indexing, query processing starts with tokenization, morphological analysis, and chunking, resulting in a query with uninflected forms, idioms and multi-word phrases identified. If a user specifies a date or a query requiring an exact match, then the date normalizer or phrase matcher is called in the appropriate case. All other queries are further processed by the query expansion modules.

The query expander consists of three parts. The first is the call to the semantic net (described more fully below) which expands the terms in the query with related terms. When a user is unsure about a spelling of a term, the fuzzy spell checker will use terms that may be orthographic variants to the original query. Finally, wildcards including arbitrary characters will suggest other term variants. The result of this pipeline is an expanded set of query terms derived from the terms in the user's query.

Based on the searching needs of on-line service subscribers and intelligence analysts, ConQuest has been running benchmarks on 30 and 100 gigabyte collections. Recently, they have been stress testing their system on 500 gigabytes of data.

The Semantic Network

ConQuest's semantic network is a knowledge base of over 400,000 concepts and 1.6 million word relationships (or word links) and was derived from electronic dictionaries, thesauri, glossaries, and other lexical sources including WordNet. In actuality, the network has two components, namely, a dictionary and a network of nodes.

The dictionary contains word sense profiles including word meanings, parts of speech, and root

and variant forms. Each word sense in the dictionary corresponds with a node. These word sense nodes are linked based on a set of relationships including synonymy, antonymy, related words, contrasted words, child-of, parent-of, part-of, and contains part-of. Simple examples of each are:

synonyms—"happy" is synonymous with "glad, joyful"

antonyms—"happy" is the opposite of "sad"

related words—"happy" is related to "merry, jolly"

contrasted words—"happy" contrasts with "woeful, sorrowful"

child-of—"sphere" is a child of the concept "geometric volumes"

parent-of—"sphere" is the parent of the concept "globe, soccer ball"

part-of—"foundation" is a part of the concept "construction"

contains part-of—"automobile" contains parts of the concepts "accelerator" and "brake"

Each type of relationship has an associated weight which represents how close the meanings of a pair of concepts are. A synonym link, for example, will have a higher weight than an antonym link. These weights are user assigned.

ConQuest provides a number of vertical thesauri, dictionaries, and glossaries including those based on the Unified Medical Language System meta-thesaurus of the National Institutes of Health and the Defense Technical Information Center thesaurus. In practice, dictionaries can be layered with up to eight dictionaries, each of which can contain over half a million word meanings. Thus, there could be a general English dictionary, a vertical dictionary of specialized terminology, and one or more enterprise, workgroup, and user dictionaries. Any or all of the dictionaries can be searched in a session.

Query Modes

The query process is done in one of three ways in ConQuest. The first and simplest is with keywords and Boolean operators AND, OR, NOT, and WITHIN. WITHIN is a proximity operator which is used with a numerical designation for the number of words within which two query terms occur. (ConQuest also offers ADJ, denoting a bi-directional adjacency operator.) Parentheses may be used to bracket particular phrases within a Boolean query.

While the standard Boolean mode searches for exact matches of terms, it can be extended to retrieve inflected forms of terms. The query words designated for morphological expansion are specified by the user.

The second search method is called the Smart mode, which has nine levels of query expansion, namely, exact matches and morphological variants, inflected forms, neighboring terms, strong synonyms, synonyms, strong antonyms, antonyms, related words, and contrasted words. The distinction, though, between a synonym and a strong synonym or an antonym and a strong synonym is subtle. For instance, the word "ball" in its sense of dance has "cotillion" as a strong synonym and "prom" as a weak synonym. Although several of the expansion capabilities map directly onto the semantic links for dictionary entries (described above), the full specification of the relationship between the links and expansion levels is proprietary to ConQuest.

In each of the expansion levels, users can adjust the weights to assign importance to a particular

type of expansion at the cost of the others.

The third mode is called the Expert mode and it allows users to restrict searches to particular senses or meanings of terms in a query. For instance, in a case where a query has a term with multiple meanings, the user can select which sense (or senses) of the term should be used in the search. For instance, in the query "bank collapse," the user can specify that searching should be constrained to the sense of "bank" related to financial institution and not river bank.

Functionality of the Smart mode is included in the Expert mode as is the capability for modifying weights assigned to search term senses.

Other Search Capabilities

ConQuest supports:

- Fielded queries, allowing users to search for authors, titles, or other information in fields.
- Recurrent queries, where searches are restricted to the results list of a previous search.
- Query by example, where a relevant document is used as a query.
- Fuzzy spelling, permitting users to search for terms with variant spellings.
- Date/number ranges, allowing users to search for dates and numerical ranges (10-100 trillion, > 100 yen).
- Wildcards, enabling users to insert a wild card at the middle or end of a word to expand the variant forms of the word.
- Re-using queries, where up to 25 queries can be saved and re-used in a session.

Differentiating Features

The key feature of ConQuest's technology is its large dictionary and semantic network that permits query expansion during searching. While other companies offer dictionaries, thesauri, and APIs to external lexical components, none has the size that comes bundled with ConQuest's software. With the presence of the dictionary and associated semantic network, ConQuest is able to offer three modes of searching, two of which (Smart and Expert modes) enable query expansion based on word senses.

Miscellaneous

According to the marketing literature, the roots of ConQuest are not based in natural language or artificial intelligence, but in radar engineering. Paul Nelson, who is a co-founder of ConQuest, developed radar engineering software at Westinghouse. The software was designed to extract

information about targets from clutter and noise signals in radar maps and was then used to build statistical models of expected targets.

While working on the radar engineering software, Nelson took a natural language class at Johns Hopkins taught by Ed Addison, a former colleague of Nelson's at Westinghouse and at the time the director of the artificial intelligence lab at Booz, Allen & Hamilton. As a result of this association and their current work, the two came onto the idea of applying the radar metaphor to text retrieval. Although their approach does not imply that they eschew natural language processing or artificial intelligence, Addison and Nelson believed that radar engineering gave a valuable perspective on improving information retrieval. With funding from the U.S. Air Force, they formed a company to develop search and retrieval technology based on their novel approach.

ConQuest has several patents pending, including some for the idea and technique of using published dictionaries as a basis for concept-based retrieval through word meaning.

Cuadra Associates, Inc.

Headquarters

Cuadra Associates, Inc.
11835 W. Olympic Blvd., Suite 855
Los Angeles, CA 90064

Tel: 310.478.0066 or 800.366.1390

Fax: 310.477.1078

email: sales@cuadra.com

WWW: <http://www.cuadra.com>

Corporate Fact Sheet

Business

Cuadra Associates develops and markets information retrieval products.

Cuadra was founded in 1978 and its president, Carlos A. Cuadra, headed the information research department at System Development Corporation (SDC). He was appointed by three U.S. presidents to successive terms on the National Commission on Libraries and Information Science.

Cuadra's programmers were, at SDC, the original designers and developers of the National Library of Medicine's MEDLARS information retrieval system and ELHILL on-line retrieval system. Both systems have been running since 1970.

Cuadra's main product, STAR, has been in the marketplace since 1982.

Markets

Cuadra's main markets are records management, libraries, archives and museums, indexing and abstracting services, and database services. Although records management is Cuadra's fastest growing market segment, libraries comprise approximately half of Cuadra's customer base.

Sample Applications

National Agricultural Library—National Agricultural Library uses STAR for several bibliographic databases, scientific research databases, mailing lists, and databases from which CD-ROM and print publications are generated.

Royal Armouries at HM Tower of London—Cuadra’s information retrieval technology is used to search records for all the collection’s objects.

Oxford University Press (Australia, Canada, UK, and U.S.)—Oxford University Press has STAR databases of bibliographic records, library records, customer service records, and mailing lists.

Other STAR users include Library of Congress (several divisions), Federal Deposit Insurance Corporation, Food and Drug Administration, Patent and Trademark Office, American Institute of Certified Public Accountants, Bank of Boston, Hughes Aircraft Co., MCI Telecommunications Corporation, National Association of Insurance Commissioners, Shell Oil, SRI International, and Syntex.

Products

Core Technology

STAR is a multi-user text management system for document and image retrieval. STAR is highly customizable and gives non-programmers the capabilities for constructing a new database, adding or editing database fields, changing the search interface, or generating a new report format.

Related Products

Cuadra offers several text and image management products including:

STAR/Client is a Microsoft Windows-based application to support searching and report generation by PC users in STAR databases.

STAR/Web provides a common gateway interface for World Wide Web servers to support STAR searching and report generation and forms-based input by standard Web browsers, e.g., Mosaic and Netscape.

STAR/Libraries is a product for managing traditional library materials (e.g., monographs and serials, reports, patents, and audio-visual materials) and full-text documents.

STAR/Rims is a records management product. It can be use to create and maintain inventory data, maintain retention schedules, link retention schedules to inventories, maintain statistics of services, and generate reports and on-line displays.

STAR/WorkSaver helps to manage documents created through word processing in departments within an organization. It can be used to capture documents when they are created, and index the documents automatically.

STAR/Images is an integrated product of STAR and Co-Star. Co-Star is an image management system developed by Alpha Pacific. It supports scanning of multi-page black-and-white

documents into collated documents, importing and collating of previously scanned documents, multiple storage media on the image server, image retrieval, and printing and faxing.

STAR/Thesaurus Application Package supports the development and maintenance of thesaurus databases (following the guidelines in ANSI/NISO Z39.19-1993). (STAR does not include a thesaurus.) The package supports the creation and deletion of reciprocal entries in a thesaurus, including cross-references between non-preferred terms and one or more preferred terms, broader and narrower terms, and related terms. There is a management function whereby information such as the source of a term, date added, history notes, and user/indexer scope notes are stored. Several different types of reports can be generated, including the contents of the thesaurus listed alphabetically, the history of when terms were added, deleted, or modified, hierarchical relationships between terms, and statistical information with counts of terms with various numbers of broader, narrower, and related terms.

Platforms

STAR runs on UNIX servers including x86 Intel-based PCs, under SCO Unix, IBM RS6000 under AIX, HP-9000 RISC systems under HP/UX, and Sun SPARCstations and servers under Sun OS and Solaris.

Technical Fact Sheet

Indexing

Inverted files are used for indexing. Currently, there is no list of stop words and each word in a document is indexed. However, through functionality in the API, user-specified words can be excluded from the retrieval process.

Although the number of records that can be supported depends on the amount of disk storage, the theoretical limit is 268 million records.

Records and fields within records can be of variable length. Each record can contain up to 32,000 lines of text and 500 different fields of information.

STAR applications typically have over 1,000,000 record databases and some applications have about 3,000,000 records. Although records are usually a few thousand words, some databases have several hundred pages of text in a single record.

Searching

Searching features include:

Boolean operators—STAR supports the Boolean operators AND, OR, and NOT.

Query Expansion—A thesaurus or dictionary can be used to expand the set of terms in a query.

Search numbers and editing—Each search expression is associated with a system-assigned search number that can itself be used as a search term.

Multi-field searching—Searches can be conducted on several fields with the same query.

Proximity operators—Searches can be restricted to ordered adjacent terms, terms within a specified distance from one another, or terms within a sentence, a paragraph, or the same field.

Wildcards—Terms can be truncated using trailing or embedded wildcards.

Search Tracing—When searching with wildcards or on multiple fields, STAR's trace facility will display each constituent term and the associated number of hits for each term.

Numeric Searching—Queries that involve numeric values and ranges (i.e., values between two numbers) are supported.

Hits Analysis—Users can specify the number of hits to limit retrieval.

Saved Searches—Users can save queries for subsequent searches.

STAR also supports several report generation facilities for displaying, writing (to files), and printing reports based on the results of searching.

Differentiating Features

Cuadra Associates offer a wide range of products that go beyond information retrieval technology. A large number of Cuadra's customers are libraries, while records management applications are growing quickly. For each of these types of customers, Cuadra offers a different software package, namely, STAR/Libraries and STAR/Rims, based on its core technology, STAR.

Cuadra also offers a product (STAR/Thesaurus Application Package) for constructing and maintaining thesauri.

One aspect that Cuadra features is its ease of use. STAR is reported to be modifiable (e.g., database construction and modifications, search screen development, and report generation) by users who are non-programmers.

Data Retrieval Corporation

Headquarters

Data Retrieval Corporation
11801 W. Silver Spring Drive
Milwaukee, Wisconsin 53225-1984

Tel: 800.421.3282

Fax: 414.536.1984

email: Not available.

WWW: <http://www.dataret.com>

Corporate Fact Sheet

Business

Data Retrieval Corporation develops and markets text management software. It also provides consulting services to organizations in the public and private sector.

Data Retrieval was purchased by and is a subsidiary of West Publishing.

Data Retrieval was founded in 1964.

Markets

Data Retrieval focuses on three vertical markets, namely, financial institutions including banks, insurance companies, and state governments. Specific applications that Data Retrieval specializes in are retrieval technology for on-line manuals and document assembly and authoring.

Sample Applications

Provident Life and Accident (Chattanooga, TN)—Provident uses TextBOOK for searching and managing compliance bulletins.

Recreational Equipment, Inc. (REI)—REI uses TextBOOK as a replacement for its three-volume, 500-page human resources manual. REI's Human Resources, Management Information Systems, and Public Relations/Public Affairs have incorporated TextBOOK in their operations.

Royal Canadian Mounted Police (RCMP)—TextDBMS was used to cut costs and streamline operations by reducing paper flow. There are about 5,000 different forms used by 25,000 people at RCMP which are now processed electronically.

Other customers of Data Retrieval include John Alden Life Insurance of Miami, Florida; Signet Bank of Richmond, Virginia; and the State of Idaho.

Products

The main product of Data Retrieval, called TextDBMS Series, is a suite of tools consisting of TextBOOK, TextGEN, TextCOMPLY, and TextDBMS.

TextBOOK—TextBOOK is an application for disseminating electronic documents. It includes a search and retrieval engine. It includes functionality for adding notes, book marks, and hypertext links to other documents, forms, graphics, OLE objects, and Windows applications.

TextGEN—TextGEN supports the creation of documents. It allows information from a variety of sources to be merged, integrated, and properly formatted into a finished document.

TextCOMPLY—TextCOMPLY is a tool for managing and tracking documents in an organization. TextCOMPLY also offers security functionality such as limiting document access.

TextDBMS—TextDBMS allows integration from existing files such as databases or word processing in support of document production.

Platforms

TextDBMS runs on IBM (CICS), OS2, DEC Vax, and Microsoft Windows. A Windows NT version will be out next year. A UNIX version will be out in late 1996 or early 1997.

Technical Fact Sheet

Indexing

Inverted files are used for indexing and benchmarks are unavailable.

Searching

Search capabilities include Boolean, adjacency, proximity, wildcard, and phrase searches.

Although a thesaurus is not included with TextDBMS, query expansion is supported with user-supplied thesauri.

Stemming is unsupported.

Searches can be performed on profiles or structured information such as author or dates of documents.

Dataflight Software, Inc.

Headquarters

Dataflight Software, Inc.
2337 Roscomare Road, Suite 11
Los Angeles, CA 90077

Tel: 310.471.3414 or 800.421.TEXT
Fax: 310.471.5294
email: support@dataflight.com
WWW: <http://www.dataflight.com>

Corporate Fact Sheet

Business

Dataflight Software develops and markets information retrieval technology.

Markets

Dataflight targets a wide range of applications including board of directors' meeting minutes, field report analysis, legislation tracking, patent information, alumni tracking, and benefits planning and administration.

Products

Core Technology

Concordance—Standard Edition—Standard Edition of Concordance is a search and retrieval database system.

Platform

The Standard Edition runs under Microsoft DOS, Windows, and IBM OS/2. An Internet Web Server (CGI) version is also available.

Related Products

Concordance—The Professional Edition—The Professional Edition has all the features of the Standard Edition together with an applications development programming language.

Concordance—The Network Systems Edition—All editions of Concordance are available in

multi-user, network versions. Capabilities include support for file and record locking, simultaneous multi-user editing, searching, and reporting.

Concordance—The Runtime Module—The Runtime Module, which is designed to be distributed with applications using Concordance, restricts user access to those features built into an application.

Concordance—The Publisher's Edition—The Publisher's Edition is a read-only version of the Runtime Module.

Benchmarks

Dataflight has performed benchmarks with a number of PC configurations. The test was on a 76,000 record database that consumes over 100MB of source text. It included foreign language documents, and many numbers and unique words. The database was located on a Novell file server with 16MB RAM. The workstation and server were 90MHz Pentiums. For an index cache size of 1,572,864K, indexing took over 14 hours, whereas for an index cache of 10,690,560K, indexing took about 8 hours.

Databases can contain over 2,000,000 records. However, Concordance can search 16 databases simultaneously, so the actual limit is 32,000,000 records.

Technical Fact Sheet

Concordance search features include:

Indexing

Inverted files are used as an index. A list of stop words, which can be edited, is included.

Searching

Boolean, proximity, and wildcard (single and multiple characters) searches are supported.

Database style of queries, e.g., searches for numerical information involving equality, greater than, and less than relations, as well as searches on dates, are permitted.

Fuzzy searching including phonetic ("Jim" and "gym") and orthographic ("FBI" vs. "F.B.I.") variants are supported.

Queries can be saved and re-used in subsequent queries.

Results are returned in a list ranked by the number of occurrences of search terms in the document.

Pull-down menus assist users, and on-line context help is provided.

A thesaurus search mechanism is also available.

Databases can contain fixed length text, date, numeric, and free-text fields.

Dataware Technologies, Inc.

Headquarters

Dataware Technologies, Inc.
222 Third Street, Suite 3300
Cambridge, MA 02142

Tel: 617.621.0820

Fax: 617.621.0307

email: info@dataware.com

WWW: <http://www.dataware.com>

Corporate Fact Sheet

Business

Dataware Technologies is an international developer and marketer of software and services for electronic information providers to manage and distribute information on CD-ROM and online.

Dataware Technologies acquired BRS Software Products and its full-text retrieval software, BRS/SEARCH in March 1994.

Dataware was founded in 1988.

Markets

Dataware Technologies markets its products and services to corporations, publishers, government agencies, and universities. In particular, Dataware provides software and services for on-line information management and CD-ROM publishing.

Dataware software is used in more than 2,000 organizations and is accessed by more than 20,000 end users in 100 countries.

Sample Applications

Dataware's search and retrieval products are being used in a wide range of applications, such as for obtaining market intelligence and litigation discovery support in industry, and by the government for fraud investigation, visa application processing, and classified intelligence work.

Products

While Dataware offers a number of authoring and development software in its CD-ROM line, it has two products for text and document management, namely, BRS/SEARCH and Total Recall. With the exception of a brief description of Dataware's other products, this profile is restricted to BRS/SEARCH.

Core Technology

BRS/SEARCH is a full text information retrieval system. It offers portability across PCs, minicomputers, mainframes, and client/server environments. BRS/SEARCH includes a report writer for customized screen and print formats of retrieved information, security capabilities for restricting user accessibility to documents or parts of documents, and multimedia and image support. This last feature enables users to create links between text, images, sounds, and video on hard disk, CD-ROM, and other media.

Related Products

Products related to BRS/Search

BRS/Word Plus automates the loading of word processing documents into BRS/SEARCH while maintaining their original markups.

BRS/Image converts images, facsimile files, and any TIFF or CCITT Group IV files for storage and retrieval.

BRS/SWIFT is an application development toolkit to build BRS/SEARCH interfaces in the MS-Windows environment.

BRS/Thesaurus supports the construction and maintenance of thesauri. Related words and phrases, synonyms, typographical variants, abbreviations, and other relations are supported. After setting up the lead term in a thesaurus, a user specifies related terms (following ANSI standards). BRS/SEARCH automatically generates reciprocal relationships between the same terms. BRS/SEARCH can also be used to verify searches for incorrect or non-preferred terms.

LEXSEARCH software allows users to search databases using the same commands as found on the LEXIS service.

The Dataware Natural Language Object Library is an API for developing natural language capabilities for applications using BRS/SEARCH. The Natural Language Object Library includes a sentence analyzer, a part-of-speech tagger and noun phrase extractor, morphological analyzer (inflectional and derivational), and a part-of-speech disambiguator. The library comes with a lexicon of word roots and inflections. The Natural Language Object Library supports English, French, and German.

The Semantic Network Object Library is an API for developers to produce tools for constructing and customizing thesauri or semantic networks. The Object Library also supports access to

hierarchical and non-hierarchical thesauri (ANSI Z39.19-1980). The Semantic Network Object Library supports eleven different languages, namely, Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Spanish, and Swedish.

Some of the core technology underlying the Natural Language and Semantic Object Libraries was licensed by Dataware from INSO Corporation. The technology has been extended and customized by Dataware.

Total Recall

Total Recall is an application programming interface (API) for BRS/SEARCH which enables developers to integrate and develop a common interface to text, structured data, and multimedia through connecting BRS/SEARCH to any of the major RDBMS and 4GLs.

CD-ROM Products/Services

CD Author includes CD Author Development System, CD Answer Retrieval Software, and Advanced Design Library. CD Author Development System is a menu-driven tool for guiding users in developing CD-ROM titles. The development system supports fielded, variable length data, input of unstructured data, a variety of indexing options, and international language capabilities (British English, Canadian French, Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Japanese, Norwegian, Portuguese, Spanish, and Swedish).

CD Answer is the retrieval software for databases created using CD Author Development System. Searching capabilities in CD Answer include numeric comparisons, wildcards, adjacency, phonetic, Boolean, and index browsing and selection.

The CD Author Advanced Design Library is a developer's toolkit for creating custom applications.

Reference Set

Reference Set is a set of software products for producing CD-ROM publications such as technical manuals, encyclopedias, catalogs, and policy and procedure manuals.

CD-ROM Preparation Systems

CD-ROM Preparation System is Dataware's CD-recordable software, used in the pre-mastering of CD-ROM titles.

Online Distribution Systems

Dataware Internet Server supports publishing and searching on the internet. The WWW server (residing on the Internet side) handles WWW browsing requests directly. It passes all requests for database searches on to a standard interface layer supplied with the Dataware Internet Server. This layer translates each search request into the appropriate search engine API and invokes the corresponding Dataware search engine.

The search engine delivers retrieved data to the interface layer, which formats it according to HTML rules and returns it to the server. That server, in turn, sends it out over the Internet to a Web browser.

Search capabilities for the Internet Server include multi-field searches, keyword and phrasal queries, thesaurus support, index/dictionary browsing, logical adjacency (AND, OR, NOT, and WITHOUT), and query refinement where queries can be modified without retyping the entire query.

In addition to its CD-ROM products, Dataware offers a number of CD-ROM-related services including title development, data and text conversion, project management and consulting, CD-ROM mastering and replication, CD-ROM drives and maintenance, database development, interface design, and systems integration.

Platforms (only for BRS/SEARCH)

Mainframes

Cray YMP/UNICOS
IBM MVS/CICS
IBM VM/CMS
Tandem Guardian

Workstations

DEC Alpha Open VMS
DEC VAX Open VMS
Hewlett Packard HP/UX
IBM RS/6000 AIX
ICL DN/X
Sequent/Unisys PTX
Sun Microsystems Solaris x.xxx

Deskside systems

BULL BOS
Intel UNIX System V.4
MS-Windows NT

Client/Server

Server operating systems
Sun Version 4.1.3
SCO UNIX
IBM RS/6000 AIX 3.2
HP 9000/7xx HP-UX
Sun Microsystems Solaris 2.3

Networking

Novell LAN Workplace for MS-DOS
Wollongong's Pathway Access
PC/TCP from FTP Software

PC-NFS from Sunselect
Winsock 1.1
Client operating systems
MS-DOS
MS-Windows
UNIX (Motif)

The Natural Language Object Library requires Borland C++ 3.1 or Microsoft C/C++ version 7.0, and is available under Microsoft Windows and Windows NT. Languages supported are English, French, and German.

The Semantic Network Object Library requires Borland C++ or Microsoft C/C++ version 7.0. It also requires CD Author Development System version 3.21 to load developer-defined thesaurus or semantic network terms. The Semantic Network Object Library is available under DOS and Windows NT.

Technical Fact Sheet

Indexing

BRS/SEARCH uses an inverted file structure which stores the position of every word in each document in the databases. A proprietary compression algorithm is used for the text and the inverted files. (Compression of the text yields a 30% to 50% savings.)

Databases may be made up of 16 parts, each of which may contain over 16 million documents. Each document may contain up to 65,000 paragraphs of 255 sentences of 255 words each. (The limitation on sentences and words are due to internal constraints on indexing only, not constraints on documents displayed.)

There is no inherent limit on the number of databases beyond hardware and operating system restrictions.

Fixed length fields (e.g., a date field) can be defined. Additions or edits to these fields are automatically verified.

Any document can contain up to 65,000 fields and each field can contain up to 65,000 characters.

A list of non-searchable stopwords are included with BRS/SEARCH. However, users can create their own lists as well as an abbreviations list.

Non-searchable punctuation and other characters are not placed into the inverted file. Users do have an option for determining how hyphenated, possessive, and abbreviated words are handled.

BRS/SEARCH includes a facility for reporting distributional information such as number of searchable words, paragraphs, and documents.

Retrieval

BRS/SEARCH supports the following search capabilities.

Relevance ranking—Relevant documents are sorted and displayed according to the number of occurrences of query terms.

Saved searches—Users can save searches (permanently or temporarily) for use in subsequent searches.

Logical searching—Searches can include the operators AND, OR (inclusive OR), XOR (exclusive OR), and NOT.

Positional searching—Search terms can be restricted to the same paragraph or same sentence. Directional and bi-directional operators are supported and can be used in combination with the operator restricting searches to the same paragraph.

Numeric searches—Numeric fields can be searched using the relations greater than, less than, and equality.

Nested searches—Queries with nested expressions are supported. There is a limit of 15 levels of embeddings.

Truncated word searches—Users can use wildcards (anywhere in a search term) as part of a query.

Saved searches—Searches can be saved and re-used or edited as parts of a new query.

Singular/plural searches—As an option, singular and plural forms of the search term are retrieved automatically and merged together as a single result.

BRS/SEARCH offers an additional module to create and maintain an ANSI thesaurus. Relationships that can be defined between terms are lead term, related term, broader term, narrower term, preferred term, non-preferred term, synonymous term, scope note, language term, date/history note, and other term (user-defined).

BRS/SEARCH provides three commands for displaying each word in a database (together with the number of documents containing each word and the number or word occurrences). These commands are ROOT for words that begin with a particular root, PREF for words that end with a certain stem, and EXPAND for displaying a number of words alphabetically preceding and following a particular word.

Searches up to 255 characters are permitted.

Advanced Searching Features

In addition to BRS/SEARCH, Dataware offers a Natural Language Object Library and Semantic Network Object Library for providing advanced search facilities using natural language techniques.

Differentiating Features

Dataware produces a wide variety of products for electronic information providers ranging from CD-ROM preparation and authoring, search and retrieval technology, and on-line distribution systems. Moreover, Dataware's technology runs on many different platforms and is multi-lingual. Through its object libraries, Dataware also offers a spectrum of natural language processing modules to complement its searching capabilities.

Miscellaneous

In May 1995, Dataware Technologies was selected one of *Inc. Magazine's* 100 Fastest-Growing Small Public Companies for the second consecutive year.

Excalibur Technologies Corporation

Headquarters

Excalibur Technologies Corporation
9255 Towne Centre Drive, 9th Floor
San Diego, CA 92121

Tel: 619.625.7900, 800.788.7758

Fax: 619.625.7902

email: info@excalib.com

WWW: <http://www.excalib.com>

Corporate Fact Sheet

Business

Excalibur Technologies Corporation develops and markets indexing and retrieval software.

The kernel of Excalibur's technology is called Adaptive Pattern Recognition Processing (APRP), which allows products built on top of it to find patterns in numerous domains including text, fingerprints, still images, full-motion video, faces, voices, and signal data. APRP is claimed to be fault tolerant to errors from input, OCR, and content for text retrieval applications.

The company was founded in 1980 by James Dowe III, the developer of the technology. From 1985 to 1989, Excalibur developed its text retrieval technology, and in 1989 released its first products in this area. In 1991, Excalibur introduced PixTex/EFS (now called Excalibur EFS) for capturing, indexing, and retrieving text and images.

Excalibur Technologies purchased ConQuest Software in a stock deal that closed on July 20, 1995. It is expected that a combined offering of products from ConQuest and Excalibur will be available in mid-1996. Users will be able to search images, audio, video, drawings, or text using either the bit-level pattern recognition engine from Excalibur or semantic networks from ConQuest.

Markets

Because of the diverse applicability of its technology, Excalibur markets extend beyond those for information retrieval. At present, Excalibur has more than 20,000 users at 500 sites. Typical customers include organizations in office, laboratory, engineering, and factory environments in such markets as government, legal, manufacturing, pharmaceutical, insurance, and financial.

Sample Applications

Pillsbury, Madison & Sutro—Pillsbury, Madison & Sutro is a law firm specializing in civil and criminal litigation and has more than 100 partners and 220 associates. The firm uses Excalibur's electronic filing software for document management, i.e., for locating scanned documents including law documents, depositions, correspondences, text reports, legal memoranda, and contracts.

Solar Turbines, Inc.—Solar Turbines, Inc., a subsidiary of the Caterpillar Corporation, designs, manufactures, installs, and maintains small to mid-size gas turbine engines. More than 75% of its business involves exporting. Since international regulations require manufacturing businesses to comply with the ISO 9000 quality standard, specification updates which were paper documents had to be validated by ISO audits. Solar Turbines selected Excalibur to provide access to scanned engineering specifications and documents. Information can now be retrieved in seconds rather than up to two weeks previously.

Other customers of Excalibur include Bellcore, The Boeing Company, Chevron, Chrysler, Digital Equipment Corporation, Federal Reserve Board, Goldman, Sachs, IBM, Library of Congress, The Mitre Corporation, TRW, University of Massachusetts, and the World Bank.

Products

Core Technology

The core technology of Excalibur's products is APRP. This proprietary technology is based on a neural network model that recognizes the presence of patterns in digital data. APRP is a collection of guidelines and an architecture that describe the policies employed to generate a neural network based on incoming data.

Related Technology

With APRP as a foundation, Excalibur offers a family of products called Excalibur Recognition Software (XRS) including Excalibur TRL Text Retrieval Library, Excalibur TRS Text Retrieval Server, and Excalibur EFS Electronic Filing Software.

Excalibur TRL Text Retrieval Library is a C-callable library that provides indexing and searching capabilities. TRL supports controlled vocabulary, keyword lists, and stopword lists during indexing.

Excalibur TRS Text Retrieval Server provides distributed, multi-user indexing and searching in client-server environments. TRS uses the same C-libraries incorporated in TRL.

Excalibur EFS Electronic Filing Software is an off-the-shelf document imaging software product featuring: fuzzy searching, automatic indexing, and support for multiple databases

including Sybase, Oracle, Informix, and Rdb. It is a client/server product.

Other products in the Excalibur XRS family include Excalibur Fingerprint Retrieval Server and Excalibur XRS Image and Signal Server.

Platforms

Server Hardware and Software

ESF currently supports the following server platforms:

IBM RS6000 under AIX, Sun SPARC under Sun OS 4.1.3 or Solaris 2.x, HP900 Series 800 or Series 900 under HP/UX, DEC Vax/VMS and DEC Ultrix, DEX Alpha under OpenVMS and DEC UNIX.

TRS and TRL currently support the following server platforms:

Windows NT on RISC-based workstations, IBM RS6000 under AIX, Sun SPARC under Sun OS 4.1.3 or Solaris 2.x, HP 9000 Series 800 or Series 900 under HP/UX, DEC Alpha under DEC UNIX, Silicon Graphics under IRIX.

EFS client platforms are PCs (386 class or above) running Microsoft Windows 3.1 (or Windows for Workgroups) or Apple Macintosh using MAC OS System 7.

TRS and TRL support clients under PCs running Windows along with the above mentioned UNIX platforms running as clients in distributed systems.

Excalibur's ESF currently integrates with Oracle, Sybase, Informix, Ingres, and Rdb.

Technical Fact Sheet

Indexing

Excalibur uses a proprietary character-based vector model for indexing.

Indexing is done on full text and is automatic when a new document is added to a collection. No preprocessing is required for documents being added and indexed.

Stopword removal is available through the API in TRS.

Indexing overhead is as low as 30%.

On a SPARCstation20 with 32MB RAM, 5MB of text can be indexed in approximately 45 seconds.

Search Capabilities

APRP is the core of Excalibur's search technology. EFS supports fuzzy searching and a fileroom metaphor as a graphical user interface for searchers (both are described below).

EFS's Fileroom metaphor is an environment for hierarchical filing. The interface consists of icons of different filing cabinets that have drawers with folders. Each cabinet represent different collections that have been indexed and are ready for searching.

EFS provides attribute or key word searching by its links to relational databases, namely, Oracle, Sybase, Informix, Ingres, and RdB. Traditional database type of searching is available for searching documents based on attribute value. EFS also allows for label searching on any label in the fileroom, i.e., document name, folder name, drawer name, and/or cabinet name.

EFS Retrieval Features

EFS uses a fuzzy search method for retrieval together with several user-controlled search option, namely, keyword, phrase, rated phrase, Boolean, and thesaurus search options. TRS provides a wider range of search features than EFS.

Fuzzy Search Type—Each word in a query is expanded to a list of words that contain similar patterns (e.g., letter substitutions and transpositions). Users may specify the depth of these lists. The search process looks for matching patterns and returns relevance-ranked vectors including hit number, score, and original data link.

Boolean Search Option—The Boolean operators (AND, OR, and NOT) and compound Boolean operators can be used in queries.

Phrase Search Option—Phrases of any length can be used as queries. In a phrase search, all of the words in the query must appear in the source document, in the same order as the query and without additions or omissions.

Rated Phrase Search Option—A rated phrase search will retrieve approximations of phrases, allowing hits on phrases with missing or different words to be included in the hit list.

TRS Retrieval Features

In addition to EFS's retrieval features, TRS additionally provides keyword and proximity search options, standard and weighted term feedback search types, and fuzzy search methods.

Keyword Search Option—Queries can contain one or more keywords. All the words in the query must exist in a document to produce a maximum score.

Proximity Search Option—Searches can be restricted to a user-specified distance between terms in a query.

Standard Search Type—Standard searches seek exact matches for terms in a query. No matches will be returned if there is not an identical match.

Weighted Term Feedback Search Type—The weighted term feedback search returns lists of words containing patterns similar to each word in a query. Separate lists for each term and each index are presented. Users then may select any of the words from the list, using them to re-form their query.

Fuzzy Search Methods—There are several options for fuzzy search methods in TRS, namely, transpositions, substitutions, patterns, and rate. The default search methods are patterns and rate.

Transpositions—Documents containing words with transposed characters can be found with the transpositions option.

Substitutions—Documents containing words with a single letter substitution (e.g., "Cheryl" and "Sheryl") will be found with the substitutions option.

Patterns—This option is used for finding identical patterns of letters in words. It can be used for finding words with the same root, but different stems.

Rate—Matches are ordered using Excalibur's standard relevance rating.

Excalibur supports term highlighting which can be modified by the user to reflect exact versus fuzzy matches. That is, terms (e.g., characters) that match a query exactly or those terms containing precise matches can be highlighted.

Differentiating Features

Excalibur uses its own proprietary pattern matching algorithms for indexing and searching. It is based on a character-based neural network that identifies patterns in digital information. This technology has application outside of information retrieval and has been used in fingerprint and image analyses.

Excalibur also provides users with a range of methods for fuzzy searching including character transpositions, substitutions, and patterns.

Fulcrum Technologies, Inc.

Headquarters

Fulcrum Technologies Inc.
785 Carling Avenue
Ottawa, Canada K1S 5H4

Tel: 613.238.1761 or 1.800.FULCRUM
Fax: 613.238.7695
email: info@fulcrum.com
WWW: <http://www.fultech.com>

Corporate Fact Sheet

Business

Fulcrum develops, markets, licenses, and supports search and retrieval software. Founded in 1983, it claims to be the world's leading supplier of text retrieval technology.

Markets

Fulcrum's directs its marketing activities to medium to large companies. Fulcrum engines are embedded in products of computer manufacturers, software vendors, and VARs.

Since its launch in 1993, Fulcrum SearchServer has been licensed to over 30,000 end users in corporations and other organizations throughout North America and Europe.

Some Key Alliances and Customers

Fulcrum has relationships with:

Bell & Howell—University Microfilm (UMI), a wholly-owned subsidiary of Bell & Howell, has a multi-year agreement for using Fulcrum's technology in a number of UMI's on-line products.

Griffiths Laboratories—Griffiths Laboratories, a world-wide food-processing company, has implemented a Fulcrum-based document and image management system to increase productivity in the area of R & D and shipping.

Microsoft, Inc.—Earlier this year, Microsoft Corp. selected Fulcrum's SearchServer for use in The Microsoft Network, Microsoft's entry into on-line interactive services.

Fulcrum is also used by Frame Technology, Glaxo, Inc., IBM, Interleaf, Mobil Oil Corporation, NEC, Siemens Nixdorf, and Sun Microsystems.

Products

Core Technology

Fulcrum SearchServer

Fulcrum SearchServer is an indexing and retrieval engine for full text applications and is the core of Fulcrum's products. The query language called SearchSQL is based on SQL.

The API is based on ODBC and SearchServer tables (described below), is ODBC-compliant, and can be viewed by any ODBC-enabled application (e.g., Microsoft Query or Word).

SearchServer is designed with a client-server architecture to support distributed computing.

Related Products

Fulcrum SearchBuilder

The SearchBuilder is a set of graphical development tools for integrating SearchServer with custom applications for Windows. Environments supported are Powersoft PowerBuilder, Microsoft Visual Basic, and Visual C++.

SearchServer Software Developer's Kit (SDK)

SDK is offered to C programmers developing SearchServer-based applications on Windows and non-Windows platforms. Access to SearchServer functionality is provided through the SearchServer API. This interface consists of 43 C language routines that provide connection management, SearchSQL language statement processing, search results processing, and error processing.

Fulcrum Surfboard

Fulcrum Surfboard is Fulcrum's newest product. Based on SearchServer, Surfboard allows companies to publish information on the internet and permits World Wide Web and other Internet users to search and navigate through the publications. Existing SearchServer collections can be made available on the internet without any requirements to re-format data. Fulcrum Surfboard supports Internet browsers such as Mosaic, Netscape, WAIS, Z39.50, and Gopher, as well as clients from AOL and Delphi.

Supported Platforms

Fulcrum products are available on more than 20 different hardware platforms and operating systems, including Windows, Windows NT, a wide variety of UNIX platforms (HP, Olivetti LSX, Siemens WX, and Sun SPARC), OS/2, and Apple Macintosh.

Technical Fact Sheet

The Fulcrum SearchServer provides the search and retrieval capabilities to Fulcrum's product suite. SearchServer supports documents in their original format in their existing locations. Text stored in relational databases may also be transparently searched and retrieved. There is no requirement that the text has to be structured or in a specific format to be accessible by SearchServer.

Text and summary information associated with a document, e.g., title or author, are organized and referenced through table structures. SearchServer builds an index of terms or words, and the index files are accessed during the search process. The original documents are only accessed for display and indexing purposes.

Table Creation

SearchServer's capabilities are contingent upon the creation of tables in which text is loaded. These tables consist of rows and columns with rows corresponding to text entities which are usually documents. Each column corresponds to one attribute of the text object and references a searchable or retrievable text element of a document. Information such as system filenames and format information are contained in table columns as is information relating to document author or creation/modification dates. Although text remains in native format and location, the content of documents is accessed as though the text is a column in a table.

Any attribute such as title or author that is displayed in a search result list should be included in a table to ensure optimal result list display performance.

There is no limit on the number of tables that can be created, and the maximum number of rows in any one table is approximately 16 million.

Structured data is supported by SearchServer for any column designated as text. These zones can be searched separate from unstructured text and are useful for searching for numeric ranges.

A column can contain up to 1,000 separately identifiable zones to a maximum of 64,000 zones per table.

SearchServer also includes a set of text readers that allow applications to access text objects in different formats and permit text objects of different formats to be grouped in a single table. These text readers do not alter the location and original document sources.

Indexing

Indexing proceeds after a table is loaded with data. Each table must be indexed before searching is possible. A tokenizer uses a set of word recognition rules to identify indexable terms during the indexing process.

Each occurrence of every term in each document is recorded, with two exceptions. Portions of documents can be marked as "non-indexed" prior to indexing and these parts are ignored during indexing. Words on a stop list are not indexed. (Word stop lists are associated with each table. Although default stop lists are provided, users can customize them to meet their needs. The limit of a word stop list is 1,024 terms.)

Columns with numeric values can be designated "value-indexed." Value-indexed terms are either dates or integers and can be searched more efficiently. Numeric range searching is supported for value-indexed terms.

Alternatively, numbers can be indexed as words and searched like other terms. However, no

numeric range searching is permitted when numbers are indexed as words.

Columns can be designated for literal indexing so that any character sequence including punctuation and spaces will be indexed and retrieved as a single term.

Non-static data can be either be re-indexed immediately to reflect changes to the data or periodically in batch. While it is not recommended that periodic indexing be performed during a period of heavy system load, it is possible to index a table that is being searched or to search a table that is being indexed.

The indexing overhead typically ranges from 20-50 percent of the size of the original text.

The Search Process

Searching a repository of documents is done via Fulcrum's SearchSQL language, which is a subset of standard SQL that has been modified to enable full-text retrieval. In other words, SearchServer's search and retrieval functionality is accessed through the SearchSQL language. SearchSQL language statements are constructed within an application and passed to SearchServer via the SearchServer API.

Once a query is entered, the application generates the SearchSQL statements according to the query. SearchServer returns the hits or documents in a working table structure. The application then builds and displays the search result list.

The application provides navigation capabilities for the user to move from line to line or page to page. Once a document is selected the application permits scrolling.

Subsequent searches can be restricted to a previous result set.

Searching, updating, and indexing may all proceed simultaneously.

Search and Retrieval Features

Key features of Fulcrum SearchServer include:

Intuitive Searching

Intuitive Searching capability provides for automatic query generation where a complete document or a fragment can be used as a query. Documents determined to be similar to the query are identified and ranked as to their relevance.

Intuitive Searching also supports natural-language-like queries.

Boolean Searching

SearchServer supports the Boolean operators AND, OR, and NOT. Search terms in a Boolean expression may be individual words, word stems, numbers, dates, exact phrases, or literal strings. Searches may be sensitive or insensitive to case.

A query may contain combinations of terms and search operators. Boolean search criteria may be used with Intuitive Searching.

Documents are ranked by number of terms per document.

For natural language queries, the terms are OR'd after stop words are removed.

Fuzzy Boolean

The fuzzy Boolean capability is used primarily for the Boolean operator AND. It is a relaxation technique whereby conjoined terms of a query are automatically and incrementally replaced by the same terms but with the OR operator. For example, consider the query A AND B AND C, where A, B, and C are terms. To widen the search for these terms, the user can invoke the fuzzy Boolean search capability and the system will also search for A AND B OR C, A OR B AND C, and A OR B OR C.

Phrase Searching

Text may be searched for an exact match of an ordered set of words or phrase.

Proximity Searching

The proximity search feature retrieves documents containing two search terms when they occur within a specified distance of each other. Both bi-directional and uni-directional proximity searching is supported.

Numeric Data

Numeric data may be searched for an exact match or for a range of specified integer values.

Date Ranges

Date columns can be created and searched for an exact match or for a range of dates.

Wildcards

Search terms used with Boolean operators may contain one or more embedded single character wildcards. Wildcards can also be used for searching for inflected (i.e., linguistically-related) forms of words.

Search Term Variants

SearchServer searches for a term exactly as provided by the user. SearchServer does not do automatic stemming and singular and plural terms are not considered equivalent. Users must specify variants using Boolean operators.

SearchServer has a facility for allowing users to use a hyphenated word as a search term to search for non-hyphenated forms as well.

European language linguistic variants are supported where character set restrictions have been imposed by the hardware or software environment in which the text was created. For example, searching for a German word with an umlauted u would retrieve a term with the umlauted character spelled out as "ue."

Thesaurus

SearchServer incorporates a thesaurus facility that supports user-defined synonyms which can be words, numbers, abbreviations, or phrases. This same facility supports user-defined suffix rules, which permit searching of inflected forms of a term. The intent of this facility is not to provide a thesaurus with wide coverage of general terms but to provide a mechanism by which domain-specific knowledge can be added. Users can also use different thesauri in executing a search. For instance, in a single query, a company name thesaurus can be used for searching a company name field, while a more technical term thesaurus is used for searching free text.

(An interesting observation made by Fulcrum is that European users rely on their own customized thesauri much more than users in the United States who do relatively little customizing.)

Search Term Browsing

SearchServer has a wordwheel facility where users have access to a list of searchable terms. A user types the first few letters of a word to browse a list of terms beginning with those letters. Relevant words are then selected by the user to formulate a search.

Term Weights

Terms in a search can have application-specific assigned weights. The purpose of weights influence the relevance value of each document retrieved.

Interim Search Result

The number of hits is reported periodically in a search. The application may be configured to terminate a search after a predetermined number of documents have been returned.

Relevance Ranking

Relevance Ranking is based on a "relevance value" for each row or document returned in a search. This value can be based on the total number of words matched in a row or on a statistically-derived value of term significance. This feature can be switched off.

Search Result Ordering

The results (rows) retrieved in a search can be ordered for presentation in order by relevance, by any column in the table (except the text column), or in chronological order. The rows can be listed in ascending or descending order.

Search Term Highlighting

SearchServer supports term highlighting where search terms are highlighted so they can be easily identified.

Search Refinement

The results of previous searches can be used in Boolean relationships with new searches or in conjunction with Intuitive Searching for further refinement.

Multiple Table Searches

A single search request can search up to 48 tables in Microsoft Windows and 198 tables in all

other environments. When tables are distributed across multiple servers in a client/server networked environment, the limits on a remote search are 3,000 tables for Windows and 25,000 tables for other environments.

Differentiating Features

Since the company focuses on embedding text management software in existing applications rather than standalone systems, the interfaces to surrounding technologies are crucial. The primary feature that sets Fulcrum apart then is its reliance on industry standards. In particular, Fulcrum's use of an SQL-based query language is a differentiating feature of Fulcrum from other IR vendors.

Other distinguishing features include Fulcrum's support for ODBC and foreign language capabilities.

Miscellaneous

Fulcrum software is installed in 40 countries and in 9 languages, namely, Danish, Dutch, English, French, German, Greek, Italian, Portuguese, and Spanish.

SearchServer can handle documents with up to 16 million searchable characters. Graphics, images, voices, or other embedded sequences of binary data are not searchable and are not counted in the limit.

It has been recently announced that Fulcrum has licensed IntelliScope Search Enhancer from INSO Corporation (formerly, InfoSoft International, Inc.). IntelliScope is a multilingual tool which will extend Fulcrum's indexing and search capabilities as well as provide other language-related functionality such as morphological analysis.

HNC Software, Inc.

Headquarters

HNC Software, Inc.
5930 Cornerstone Court West
San Diego, CA 92121-3728

Tel: 619.546.8877

Fax: 619.452.6524

email:Caid@hnc.com

WWW: <http://www.hncs.com>

Corporate Fact Sheet

Business

HNC develops and markets neural network technology for applications related to high value business problems in the area of decision support.

Markets

Currently, HNC has 18 of the 20 largest banks in the U.S. as customers for HNC's FALCON, which is a real-time, transaction-based credit card fraud detection system. HNC also markets Automated Real Estate Appraisal System (AREAS) and the Aquarius Mortgage underwriting system. HNC has a division in Atlanta that uses neural networks to forecast retail sales.

Sample Applications

MatchPlus was used in the ARPA-sponsored TIPSTER Text Detection and Extraction Program where it achieved the best recall of any commercial system. Currently, MatchPlus is being used in an alpha system by the U.S. Air Force at Wright Patterson AFB for processing classified information. It is also being used in other intelligence applications.

Besides text retrieval, HNC's technology can be used for routing, document clustering (grouping documents with similar themes), and generation of automated subject indices with explanations (determining the words whose vectors are close to each document center).

Products

Core Technology

MatchPlus

MatchPlus is a retrieval and document classification system based on neural networks. According to HNC, the techniques exploited by MatchPlus (and described below) eliminates the need for thesauri, synonym lists, knowledge bases, and conceptual hierarchies.

Related Products

CONNECTIS is an indexing and routing system based on MatchPlus.

DOCUVERSE is a document visualization system.

System Requirements

While HNC claims that MatchPlus is portable across platforms, the configuration is written in C with an X Windows/MOTIF graphical user interface running on Sun workstations under SUN/OS or Solaris or Hewlett-Packard under HP-UX.

MatchPlus has been demonstrated on collections in excess of 2GB. There are no inherent limitations in processing larger collections.

The MatchPlus provides an option of using an inverted index. This index requires approximately the same disk space as the corpus being accessed. HNC estimates that an index size of approximately 1.5 GB is needed for a corpus of 1,000,000 documents.

Technology Fact Sheet

An Overview of Context Vectors

The representation scheme used by MatchPlus is central to its techniques. Words, documents, and queries are represented by structures called context vectors which encode the meaning of a word, query, or document as a high dimension, fixed-length vector consisting of real-valued numbers or components.

Elements of the context vectors (called "features") are used for classification and retrieval of documents and are automatically generated by the system. The key notion of features and context vectors is that context vectors for a document represent the meaning of that document relative to a feature set.

During system initialization, a context vector is assigned to each unique word or phrase in the system lexicon. Training then occurs on free text that results in automatically generated word

and phrase context vectors which are used to produce document and query context vectors. (As with vector models, similarity of direction is analogous to the degree of similarity of term, word, or phrase usage.)

Context vectors for documents are based on word and phrase context vectors using the weighted sum of the context vectors associated with stems and phrases in the document or query.

Searching is achieved by the identification of document context vectors that are close to the query context vector. A document's relevance to the query is determined by comparing the dot product of the query context vector with that of the document's context vector. A large dot product suggests strong relevance to the query.

Boolean operators can be used with context vectors of queries. In all cases, documents are given a context vector ranking by a dot product. Documents with Boolean matches may, at the user's option, be placed at the top of the retrieval list independent of the context vector score if the "obey match" switch is turned on. Currently, the system supports K of N matches which is a soft combination of AND and OR. For instance, a \$Match(3)[A, B, C, D, E] would identify any document that had any three of the Boolean terms A, B, C, D, E present.

Generation of Word Context Vectors

Since searching is based on the notion of closeness of document and query context vectors, and since context vectors are supposed to represent language usage, the generation process of context vectors is critical to the accuracy of searches. The central idea behind this generation or learning process is that certain pairs of words are closer than others in a probabilistic sense. That is, words that appear together frequently (i.e., in context) are closer than those that do not. MatchPlus calculates these associations and produces word context vectors in such a way that terms that are used in a similar context will have vectors that point in a similar direction.

A critical step in generating context vectors for a document and for a query is the process of determining the word context vectors for the stems of words in documents and queries. The process of stem vector training starts with a document collection as training text. MatchPlus removes the stop words before determining a context vector. (The list of stop words can be edited by users.) Furthermore, word groups like "New Mexico" and "sodium chloride" are grouped as single terms. This grouping is optional and is based on a pre-defined phrase list. MatchPlus also derives the root of each word (suffixes and plural and tense markers are removed).

Using HNC's proprietary learning algorithm, the stem context vectors are produced and stored in a context vector table.

A user can evaluate the relationships resulting from the generation of stem context vectors by examining which words are close to a selected word. This operation, called a "stem tree", is performed as follows:

1. A user selects a word and the context vector for that word is looked up in the context vector vocabulary.

2. MatchPlus calculates the dot product which represents closeness of every other word vector in the vocabulary to the selected word.
3. The resulting dot products are sorted with larger products representing closeness of terms.

Thus, for some collections, the selected term "baseball" would be close to "baseball team" and "baseball league," but further from "soccer" or "tennis."

Indexing as the Generation of Document Vectors

Documents, phrases, and words can be represented by context vectors. Document context vectors are derived as a weighted sum of the context vectors associated with words in the document. They are also normalized to prevent long documents from being favored over short ones.

The generation process of document vectors proceeds as follows. Given a document collection, stop words are removed, phrases are identified, and stemming occurs. The resulting pre-processed documents are inverted (using an inverse frequency weight) to form an inverted index, which is a separate database from the final document context vectors. (This inverted index is used for Boolean searches as described below.)

By the process of inverted index formation, the same pre-processed documents that resulted from stemming, phrase finding, and stop list removal now have inverse frequency weights associated with their terms. At this point, a weighted sum is calculated based on the context vectors of the stems in the documents. This sum is normalized and the result is a collection of document vectors for the documents.

Retrieval Process

As with the conversion of words and documents into context vectors, queries are also transformed into context vectors. The terms in the query are processed by first removing stop words, identifying phrases, and stemming. The stems are converted to stem vectors using the stem context vectors determined earlier (through the process described above). The hits for the query are found based on comparing the dot products of the query and documents. The results list is ranked according to relevance, i.e., by the magnitudes of the dot products of the query vector with each of the documents.

Differentiating Features

HNC is one of the few commercial vendors using a neural network approach to information retrieval. Based on its core technology, HNC has developed other products related to text processing, namely, an indexing and routing system and a document visualization system. Moreover, HNC has also used its proprietary neural network technology to develop and market applications in areas such as banking, real estate, and retail marketing.

Information Access Systems, Inc.

Headquarters

Information Access Systems, Inc.
3085 Bluff Street
Boulder, CO 80301

Tel: 303.442.6224

Fax: 303.442.4530

email: holsclaw@csn.net

WWW: Not available.

Corporate Fact Sheet

Business

Information Access Systems, Inc. (IAS) develops and markets text information management products based on its proprietary technology, referred to as Judgment Space (J-Space). IAS provides consulting services to assist clients in developing information system requirements, implementations, and customized applications.

Markets

IAS markets specifically to clients with large document collections, law enforcement agencies, and CD-ROM publishers.

Sample Applications

Chemical Warfare/Chemical Biological Defense Information Analysis Center (CBIAC)—The CBIAC is a Department of Defense-chartered information access center operated by Battelle Memorial Institute under the auspices of the Defense Technical Information Center. The purpose of the project was to develop a system for delivering information to federal government contractors and university researchers who constitute the chemical and biological community. Typical requests are for reference information such as material and design recommendations that would ensure equipment survivability in potentially contaminated environments.

Other clients and applications include:

Republic of Turkey—World-wide opinion monitor

Regional Bell Operating Company—FCC and MFJ libraries

King County Police Department—Serial Homicide Investigation

State Law Enforcement Agency—Serious Offense Tracking System
Tandem Computers—Customer Support Information System

Products

Intelligent Text Management System

Intelligent Text Management System (ITMS) is the foundation of the IAS's product line. The system was designed to 1) classify and retrieve text information from very large text bases, and 2) incorporate traditional databases and keyword retrieval systems. It can operate as a stand-alone system or be integrated as a utility with existing text and database management systems. The current version, Version 5.8., runs under VAX VMS, Sun, and the DEC Alpha.

Intelligent Text Distribution System

Intelligent Text Distribution System (ITDS), V 1.0, accepts a real-time text input stream, automatically classifies the documents, and then distributes appropriate documents or document segments according to the content of interest profiles. The current version runs under Sun OS and Solaris.

Application Prototyping System

Application Prototyping System (APS) is a functionally-limited ITMS. It was developed as a support product to allow clients to prototype text-intensive applications rapidly. The APS limits users to 1.5 megabytes of text per document base.

Judgment Base Development Tools

Judgment Base Development Tools are an integrated set of four programs which support the process of developing the system intelligence, referred to as the Judgment Base Module (JBM). The tool assists the user in the selection of the vocabulary terms, manages the judgment data throughout the process, and presents the user with reports which indicate the quality of the judgments.

Judgment Base Module Products

Non-Technical English Judgment Base Module, Version 3.0.

Non-Technical English (NTE) JBM is designed to handle non-technical English language documents from a non-specific domain, such as an encyclopedia or newspaper. The current version covers 150 subject fields and includes 7500 terms and synonyms, including business, government, politics, and health. The NTE JBM can be used as an add-on for derivative applications such as legislative tracking, special interest news services, or as a navigational aid for large on-line database systems.

Serious Offense Tracking System JBM Version 2.0

Serious Offense Tracking System (SOTS) JBM handles reports on Type I crimes, including homicide, sexual assault, robbery, burglary, fraud, arson, and kidnapping. The SOTS JBM allows matching of different modus operandi (MO) descriptions. A current report can be entered and matched against previously indexed reports based on subject matter similarity.

Hardware Requirements

ITMS requires at least 16MB of memory, and more may be required depending on several factors including the size of the operating system, size of the document collection, the number of users.

Versions of IAS's software runs on Sun or other SPARC-compatible hardware, DEC VAX, and MIPS.

Operating systems requirements are VAX VMS 4.6 or later, UNIX SVR4 or compatible systems.

For distributed processing, DECnet (VMS) or TCP/IP (for UNIX) is required.

IAS expects to have a Microsoft Windows 3.1 version in 1996.

Technical Fact Sheet

J-Space technology was originally developed at the Artificial Intelligence Center of the U.S. Air Force. A J-Space is an N-dimensional Euclidean space with a coordinate system in which the reference axes are interpreted as subject-matter dimensions. Words, sentences, documents, and other textual units are assigned point-locations in the space, and the projections of each point on the reference axes are interpreted as the degree of relevance of that textual unit to that subject dimension. The procedure involves: 1) selecting a number of technical expressions or salient terms from an existing thesaurus and from the text of the documents to be indexed, and 2) obtaining scaled judgements as to the degree of relevance of each of the technical expressions, terms, or phrases to each of the subdomains of the subject matter. The judgements of relevance to the terms of a given sub-domain are made by persons who are professionally competent in that sub-domain. The result is a two dimensional matrix reflecting the relevance of each term to each sub-domain, which can be interpreted as an N-dimensional Euclidean space in which is embedded a configuration of K vectors (corresponding to the K fields) extending from the origin of the space. The configuration of vectors represents the collective scope of the K fields, and reference axes of the space provide a systematic frame of reference for representing this content domain. The process results in a Judgement Space Module, which becomes the system "intelligence" of the products based on the technology.

Indexing

Both the ITMS and ITDS perform classification rather than indexing; that is, text segments are identified in terms of their subject matter dimensions. The classification module calculates a numerical value that represents the subject matter content and places the numerical value in an N-dimensional space (J-Space). The axes of a J-Space represent subject matter dimensions. Textual units such as words, phrases, and documents are assigned point-locations in the space, and projections of each point on the axes are interpreted as the degree of relevance of that unit to a particular subject. Text classification for searching or document distribution is accomplished using a vector model where document and request vectors are compared for degree of similarity or proximity, and results are ranked based on the similarity of each text segment to the request.

(The mathematics of the system and its relationship to fuzzy logic and fuzzy Boolean operators, in particular, have been explored recently at IAS and Batelle Memorial Institute.)

The classification module can be set to identify changes of subject matter within the body of the text and assign the appropriate J-Space location based on the subject content. It will, thus, create sub-documents which results in greater searching efficiency.

The classification module does not require the incorporation of automatic stemming, or morphological or syntactic analyses.

There is no theoretical limit to the size of the document collection supportable by ITMS, other than practical limitations imposed by limitations of system storage resources. Extremely large databases can be accommodated by using the system's network query processing functions, which enable the document base to be distributed to multiple host computers that can process queries in parallel and provide a single-system image appearance to the user.

Indexing overhead is a function of average document size; the requirement for indexing a document is typically 300 bytes or less per document. Therefore, if the average document size is 6K, indexing overhead is 5% or less. Use of the subdocument feature increases this overhead as a function of subdocument generation control parameters.

Indexes are typically 10-12% of the size of the document collection, depending on the size of the sub-document window.

Searching Features

ITMS integrates three retrieval strategies: formatted fields, Boolean expressions, and intelligent retrieval (J-Space). J-Space requests are submitted to the system in conversational language. A request can be words, phrases, or whole or part of documents from prior searches. Fielded data, such as author specification, and particular key words can be specified also. In the ITMS, the breadth of the search is not determined by the request itself, but by another parameter referred to as the "search radius." The user can ask the system to broaden the search by expanding the radius. To increase the likelihood that the information the user gets is useful, the user specifies more about the subject than what is sought. That is, the more the user tells the system about what is being looked for, the more accurate the system can become.

The ITMS will retrieve documents or parts of documents and list them in order of conceptual relevance to the request. It will also highlight any keyword terms that were specified in the request.

Differentiating Features

ITMS supports conversational language or entire documents as requests and returns documents or sub-documents in order of conceptual relevance. Moreover, sub-documents are classified and retrieved separately. ITMS also supports parallel request processing and distributed database access.

In ITDS, interest profiles are stated in conversational language. Distribution is then based on individual profiles that are conceptually related to the query.

Information Dimensions, Inc.

Headquarters

Information Dimensions, Inc.
5080 Tuttle Crossing Boulevard
Dublin, Ohio 43017-3569

Tel: 614.761.8083 or 1.800.DATA.MGT

Fax: 614.761.7290

email: info@idi.oclc.org

WWW: <http://www.idi.oclc.org>

Corporate Fact Sheet

Business

Information Dimensions, Inc., (IDI) develops and markets document management tools. Its software products are installed in over 2,200 locations worldwide, including nearly 50% of the FORTUNE 100 and in major government agencies around the world.

IDI evolved out of the Battelle Memorial Institute. In the early seventies, Battelle developed a full-text retrieval system, called BASIS, which was used to search and display scientific findings from researchers in the U.S. and abroad. To market the product, Battelle formed Information Dimensions, Inc., in 1986 as a wholly-owned subsidiary. IDI continued developing BASIS, and in 1989, it released BASISplus which was the first database to combine full-text retrieval with relational database management.

At present, Information Dimensions is a subsidiary of OCLC Online Computer Library Center, Inc.

Markets

IDI's customers are typically large, worldwide institutions in a wide variety of industries. IDI serves customers in government, utilities, telecommunications, aerospace, pharmaceuticals, banking, chemical, petroleum, manufacturing, food and beverage, financial services, materials, insurance, electronics, law, and publishing. Customer applications include electronic publishing, quality management, customer service, regulatory compliance, library management, litigation support, law enforcement, and legislative tracking.

Sample Applications

Brookhaven National Labs—Brookhaven National Labs uses BASISplus in several applications including Material Safety Data Sheets, Department of Energy orders, reactor safety manual

production, and environmental and safety publication tracking.

Ameritech—BASISplus is used in an employee education and information database of news briefs, internal speeches, and public documentation.

3M—BASISplus is central to 3M's Product Information Center (PIC) which receives more than 1,000 calls a day on any of more than 60,000 products. In less than two minutes, a PIC assistant connects the caller with someone who can help.

Volvo—Information on customer service, accident damage analysis, and market research is managed at Volvo with BASISplus.

Upjohn Company—Upjohn uses BASISplus to manage documentation for U.S. Food and Drug Administration approval.

Other customers include Canadian Heritage Information Network, Canadian Department of Foreign Affairs, MITRE Corporation, Parke-Davis Medical Division, Scottish Nuclear, and Thomson Professional Publishing.

IDI has marketing agreements with CAP Gemini Sogetti, Control Data Corporation, Digital, Hewlett-Packard, IBM, ICL, Siemens, Silicon Graphics, Sun Microsystems, Unisys, WANG, and Xerox.

Products

Core Technology

Information Dimensions BASIS product suite has five product groupings. These are BASISplus, TECHLIBplus, BASIS Desktop, BASIS SGMLserver, and BASIS WEBserver.

BASISplus

BASISplus is a client/server relational database system that was designed to handle text and mixed object documents. It provides full-text retrieval with search assistance as well as database capabilities such as security, backup, and recovery. There is an integrated thesaurus and SGML support.

There is an ODBC/SQL driver that provides capabilities of BASISplus database manipulation through the Open Database Connectivity (ODBC) model. With the ODBC driver, developers can incorporate many BASISplus options and implement new applications using ODBC-enabled development tools.

All BASIS products are built on the BASISplus engine.

Related Products

TECHLIBplus

TECHLIBplus is a library automation and management system developed for corporate, legal, and technical libraries. It provides patron access, cataloging, circulation and serials control, and acquisition functions.

BASIS Desktop

BASIS Desktop is a Windows application of BASISplus and a toolkit for customizing database applications in a client/server environment. Full-text retrieval, many word processors, version control, and e-mail connectivity can be integrated.

BASIS SGMLserver

BASIS SGMLserver is a set of development tools for building database applications that store, update, and manipulate Standard Generalized Markup Language (SGML) documents.

BASIS Webserver

BASIS Webserver allows customers to publish entire BASISplus databases on the World Wide Web. Key functionality includes automatic exportation of information as valid HTML documents, concurrent information updating, and search and navigation capabilities.

BASIS WEBserver is capable of managing hundreds of gigabytes of information.

Platforms

Operating environments for BASISplus are Digital UNIX (OSF1), Open VMS, HP-UX, SunOS, Sun Solaris, SVR4 MIPS (Silicon Graphics IRIX), SVR4 Intel (e.g., Novell UNIXware), IBM AIX, and IBM MVS (TSO and CICS).

Technical Fact Sheet

Indexing

BASISplus uses B-tree for its indexing representations.

A list of stop words is provided, and users can define their own on an element-by-element (field) basis.

An individual document collection can be as large as 126GB. A user can search up to 32 collections in a given session, thereby, providing an effective base of 4 terabytes of data to search and display.

Data loading and indexing have been benchmarked at 120MB per hour on mid-range UNIX servers.

Searching

Full-text retrieval with BASISplus has the following features:

Adjacency and context searching—Searches can be restricted to specific components of documents, e.g., paragraph, sentence, or number of words, and the desired order of term occurrences.

Stopword and title control—Users have control over stopwords in that they can be either included or ignored during searches. This same capability exists for searching titles.

Term selection lists—Users have the ability to browse indexes for fields within the database to look at the terms. From these lists of terms, a user can select one or more to create a query.

Relevancy ranking—The number of hits in a document determines its score. However, users can assign weights to specific terms of a query to alter the way in which the documents are returned.

Field structure support—Searching of fields is supported, as is searching of different document types and fields with a single query. A field can contain a single or different types of values, e.g., numbers, text, dates.

Soundex and plural control—BASISplus supports Soundex-encoding which allows users to search for terms that "sound like" terms in their query. For instance, with the query term "Meyer," users could retrieve documents containing either the term "Myer" or "Mire."

BASISplus can also retrieve singular and plural forms of words including irregular inflections.

Thesaurus-based concept searching—BASISplus is capable of query expansion using a thesaurus. While Houghton Mifflin's Roget's Electronic Thesaurus is included with copies of BASISplus, users can use other thesauri. (BASISplus offers support for all thesaurus relationships defined in ANSI Z39.19-1980.) Thesaurus functions include synonym identification, controlled vocabulary, concept hierarchies, ambiguous terms, and alternate-language searching.

Synonym identification finds documents that contain synonyms of users' search terms.

Controlled vocabulary—Vocabulary control is achieved via the BASIS Thesaurus. Options exist to switch terminology in the data to preferred terminology as defined by the thesaurus (e.g., change "car" to "automobile").

The system can use a thesaurus to validate the data users enter against the terms in the thesaurus.

Concept hierarchies permit users to group words or phrases into hierarchies where the levels represent narrower or broader terms. For instance, users searching for "purchase agreements" may also want documents with the narrower term "contracts."

When a search contains an ambiguous term like "produce" (make vs. food), the user is warned

and is prompted for a meaning from the entries in the thesaurus.

Acronyms and abbreviations (e.g., "UN" - "United Nations") can be added to the thesaurus for query expansion.

Alternate-language searching enables searching to be performed in one language, say French, while the documents are in English. This capability requires that equivalent terms, e.g., "house" and "maison," are specified in the thesaurus.

Terms that relate usage or suggest other terms can also be added to the thesaurus. For instance, a term usage note such as "Burkina Faso (Before 1980 USE Upper Volta)" will prompt the searcher to use the appropriate term. Similarly, the term "vacation" may be associated with a note such as "(See also holiday schedule)" to indicate to the searcher that a related term exists.

Differentiating Features

BASISplus can be integrated with relational databases and applications and is portable across a wide range of platforms. Thesaurus support is provided. BASISplus is scalable and can handle indexing and searching of large data collections.

Inso Corporation

Headquarters

Inso Corporation
31 St. James Street
Boston, MA 02116-4104

Tel: 617.753.6500

Fax: 617.753.6666

email: To individuals only (first initial last name @ inso.com)

WWW: <http://www.inso.com>

Corporate Fact Sheet

Business

Inso Corporation (formerly, InfoSoft International, Inc.) develops and markets multilingual proofing, reference, and information management tools. These tools have been licensed to over 200 OEMs.

Since 1982, Inso operated as a part of Houghton Mifflin, and in 1994, it became an independent, publicly-held company. The company changed its name from InfoSoft International to Inso in 1995.

Markets

Inso's markets are OEMs that offer language-processing capabilities in their products. In the case of Inso's information management tools, OEMs embed Inso's technology in their applications for enhancing linguistic-based indexing and retrieval.

Sample Applications and Alliances

Fulcrum Technologies, Inc. has licensed Inso's Intelliscope Search Enhancer for use in future products. The license includes language processing capabilities form Dutch, English, French, German, Italian, Portuguese, Spanish, and Swedish. Other languages will be added upon availability.

Also, Dataware Technologies, Inc. has licensed the Intelliscope technology.

Products

Inso offers a wide range of language-oriented products including information management tools, proofing tools, and reference works.

Information Management Tools

Inso's information management tools product line consists of the Intelliscope Search Enhancer. Intelliscope is a multilingual tool that provides search expansion capabilities for applications, and can be used for index preprocessing, search preprocessing, or both. For index preprocessing, Intelliscope uses a set of linguistic functions to enable applications to process text and enhance indexes with language-specific information. Similarly, Intelliscope provides a number of query expansion techniques based on the language of the query. (See Linguistic Functions below.)

Intelliscope supports ASCII, ISO, and Unicode character sets. It currently handles Dutch, English (Australian, UK, and U.S.), French, German, Italian, Portuguese (Brazilian and European), Spanish, and Swedish. Inso has no definite plans for supporting South East Asian languages.

Inso is working on benchmarks for Intelliscope.

Beyond a brief description of Inso's other products, this profile focuses solely on the Intelliscope Search Enhancer.

Proofing Tools

Currently, Inso offers four proofing tools, namely, International CorrectSpell, International Proofreader, CorrecText Grammar Correction System, and International Hyphenator.

International CorrectSpell is a multilingual spelling correction system. It supports 18 languages and their dialects, including compounding in Germanic languages, clitic processing in Romance languages, special rules in agglutinative languages (e.g., Finnish), and possible word forms in highly inflectional languages (e.g., Czech and Russian).

International Proofreader is a multilingual proofreading system. It supports three modes or levels of checking for writing errors and style problems, namely, Spelling, QuickProof, and FullProof.

CorrecText Grammar Correction System parses sentences in order to detect and correct errors in punctuation, spelling, grammar, and style. The system matches parses with a database of grammatical rules for determining possible errors. Explanatory messages associated with suggested changes are provided to the user. CorrecText applies only to English.

International Hyphenator provides hyphenation for 18 languages and dialects. The system uses a combination database lookup together with hyphenation rules for determining hyphenation points.

Inso also offers the Personal Dictionary in conjunction with their proofing tools. It enables users

to customize their systems to include frequently used words. The Person Dictionary features dynamic RAM-resident databases that store up to 6,000 words, static, disk-based databases that store up to 300,000 words, a file compression utility that permits users to create compressed personal databases from large ASCII word lists, customized correction that enables OEMs or users to specify words to be flagged and replaced by other specific words, and compatibility with other Inso proofing products.

Reference Works

Currently, Inso markets several reference works, namely, International Electronic Thesaurus, Roget's II Electronic Thesaurus, The American Heritage Dictionaries, The Columbia Encyclopedia, The Information Please Almanac, The Information Please Business Almanac, The Dictionary of Cultural Literacy, and Simpson's Contemporary Quotations.

Associated with Inso's reference works is the Intellifinder Reference Engine. Intellifinder is an information retrieval system that uses the structure of Inso's dictionaries, encyclopedias, thesauri, and other references for locating information. For example, Intellifinder provides a table of contents browsing feature where users can view the hierarchical organization of the encyclopedia and from which any relevant section can be accessed.

Intellifinder also supports query expansion in which a set of options for conceptually-related words to a query are provided to a user.

The IntelliFinder Reference Engine supports ASCII, ISO, and Unicode character sets.

Platforms

Intelliscope runs under Microsoft Windows, Windows NT, Macintosh, Solaris, and OS/2.

Technical Fact Sheet

Since Intelliscope Search Enhancer is designed to support language-based indexing and retrieval for the OEM market, it does not provide indexing and search capabilities as most IR vendors do. Instead, Inso offers a set of language-specific linguistic functions for index and search preprocessing that can be viewed as an add-on to the indexing and searching technology provided by the OEM. The linguistic utilities are language specific and are described below.

Linguistic Functions

Intelliscope's capabilities include:

Tokenization and normalization—Text is converted into tokens using language-specific rules for processing capitalization, attached punctuation, and word-internal punctuation, e.g., apostrophes, parentheses, slashes, and dashes.

Sentence boundary identification—Tokens that begin and end sentences can be identified and marked.

Clitic processing—For Romance languages, pre- and post-clitics (attached articles, pronouns, and prepositions) can be identified and removed.

Compound analysis—For Germanic languages, compound words can be decomposed into their individual components and separators can be removed (e.g., "eisenbahngesellschaft" - railway company)

Inflection and uninflection—Inflected forms of words can be generated from root forms, and root forms are derived from inflected forms.

Part of speech tagging and verification—Words can be looked up in Intelliscope databases for parts of speech. Spelling and capitalization can also be verified at look up.

Derivation expansion and reduction—Semantically-related words can be found using Intelliscope's databases. Related words may be of different syntactic categories, e.g., "hope" and "hopeful."

Spelling variant identification—Variations in spelling including dialectical variants (e.g., "ise" and "ize" in English) can be identified.

Thesaurus functionality—Words that are conceptually similar can be identified using IntelliFinder's retrieval functions and electronic thesauri.

Noun phrase analysis—Noun phrases can be identified in text and treated as elements, e.g., "Regional Bell Operating Company."

Part-of-speech disambiguation—Parts of speech of words can be assigned based on context of the word in text.

Personal Dictionary—Users can customize the system by adding words or word senses. These include proper names, technical terms, acronyms, and addresses.

Differentiating Features

Inso provides a wide range of language-related tools and references, many of which are multilingual and have language-specific capabilities. Although Inso has developed text management software, it markets its technology to OEMs and as such does not offer off-the-shelf IR products.

Miscellaneous

The linguistic modules (noun phrase extractor and part of speech disambiguation module) were developed at Inso.

Odyssey Development, Inc.

Headquarters

ISYS (Odyssey Development, Inc.)
The Denver Technological Center
8775 E. Orchard Road, #811
Greenwood Village, CO 80111

Tel: 303.689.9998

Fax: 303.689.9997

email: isys@dash.com

WWW: <http://www.isysdev.com>

Corporate Fact Sheet

Business

ISYS develops and markets information and retrieval systems. Recently, ISYS has partnered with PaperClip Imaging Software to offer a new product with text retrieval and document imaging capabilities.

The company was founded in 1988.

Markets

ISYS has over 40,000 seats installed worldwide on single workstations, LANs, and WANs. It has a large user base in the legal profession, law enforcement agencies, and the court system. ISYS products are also used in government agencies, utilities, and accounting firms.

Sample Applications

Carolina Power & Light Company (CP&L)—CP&L, which operates three nuclear power plants, uses ISYS for managing its "Final Analysis Report" and the Nuclear Regulatory Commission-issued "Technical Specifications" that set operating limits on the plant. Both of these documents are periodically updated and re-submitted as changes occur.

Mitchell, Silberberg and Knupp—Mitchell, Silberberg and Knupp is a law firm in Los Angeles that uses ISYS to index litigation depositions on current cases. The current database is 180MB with 1400 files.

Other customers include Boeing Corporation, Nikon, Rockwell International, and the World Bank.

Products

While Odyssey Development offers publishing software (ISYS Electronic Publisher), an integrated document retrieval and imaging product (Image Suite), and a document retrieval and storage management system (ISYS and Avail NetSpace), Odyssey's core product is ISYS, its search and retrieval engine.

Platforms

ISYS runs under DOS, Windows, and Pen.

Technical Fact Sheet

Indexing

Inverted files are used in ISYS. ISYS supplies a list of 317 stop words, which can be edited by users.

ISYS can process 2 billion words and more than 65,000 paragraphs in a document. An index can contain as many as 1 million files and 2 billion words. Indexes can be compressed (with PKZip) to less than 25% of their original size. Documents can also be indexed in the Zip format and unzipped during retrieval.

Searching

ISYS supports menu-assisted and natural language searches where words or phrases can be used in a query and all must be matched.

Other searching features includes:

Truncation—leading and ending wild cards are allowed.

Date recognition—multiple date formats are recognized.

Boolean operators—AND, OR, and NOT can be used to form complex queries.

Proximity searches—uni- and bi-directional searching are permitted.

Sounds like—Queries can be expanded where similar sounding (phonetically related) terms are substituted for original query terms.

Synonym rings—Users can define synonym lists whose elements may be used in searching.

Fuzzy searching—Matches based on character substitutions (e.g., "duck" and "buck") are supported.

Relevance ranking—documents returned are ranked based on a ratio of hits to document size.

Queries can be saved and re-used in subsequent searches.

A table-of-contents outline for indexed files can be generated and used to group files by categories.

Miscellaneous

In a couple of recent product comparisons (*InfoWorld*, May 24, 1993 and *InfoWorld*, April 17, 1995) ISYS was awarded the highest score.

Open Text Corporation

Headquarters

Open Text Corporation
180 Columbia Street West
Waterloo, Ontario
Canada N2L 3L3

Tel: 519.888.7111

Fax: 519.888.0677

email: info@opentext.com

WWW: <http://www.opentext.com>

Corporate Fact Sheet

Business

Open Text Corporation develops and markets text database management systems. It provides training and consulting services for its customers. Founded in 1991, the company's technology is based on work developed at the University of Waterloo. Open Text has over 2,500 users worldwide.

Open Text claims that its technology, which can process multi-gigabyte databases, has the fastest search capability available.

Markets

Open Text has targeted the following major markets: automotive, education, financial services, government agencies, insurance, libraries, manufacturing, and publishing.

Sample Applications

Blue Cross / Blue Shield of Oregon—Open Text was selected by Blue Cross / Blue Shield of Oregon for its search and retrieval capabilities in an on-line reference manual system. Once fully deployed, the new system will replace the paper manuals that include information on customer service, claims processing, medical review, underwriting/actuarial, and provider affairs.

Continuing Education of the Bar, California (CEB)—CEB, which is one of the country's oldest and largest organizations dedicated to the continuing education of lawyers, publishes 120 titles that are updated annually. These publications are developed for programs offered by four separate areas of the company's business (books, reporters, action guides, and courses). To streamline and integrate its programs, CEB undertook a major computerization project of converting legacy text databases to a client/server environment. Open Text is providing the

technology to convert publications into electronic formats including SGML. Using Open Text, CEB is developing a full scale repository to allow customers on-line access to all CEB documents.

Legislative Assembly of the province of Ontario, Canada—Open Text provided the expertise and technology to convert the Hansard, the transcripts of speeches of the legislative body, into structured format to assist in indexing. With Open Text's retrieval engine, the Assembly support staff accesses speeches pertaining to a particular member of the Assembly during a date range or on a specific subject.

Oracle Corporation—Oracle has recently licensed Open Text's indexing and retrieval technology to be incorporated in Oracle products. It is the first non-Oracle information search technology offered by Oracle, and will complement Oracle Book's navigation and browsing facilities and other aids generated by Oracle ConText, the linguistic analysis technology. Oracle Book supports the creation and electronic distribution of corporate documents including technical manuals, corporate policies and procedures, documentation, and catalogs.

Other customers of Open Text include: Bibliotheque de France, Caterpillar, General Dynamics, Grolier Publications, Microsoft, National Security Agency, Union Bank of Switzerland, and the University of Virginia.

Products

Open Text 5 is a suite of text retrieval software tools. It consists of a search engine with Graphical User Interface (GUI) query and viewing tools, a parallel execution monitor, a text display/browser, and support for more than 40 native file formats, SGML, and HTML. These capabilities are bundled as three integrated technologies, namely, Open TextSearch, Open TextQuery, and Open TextView.

Open TextSearch includes Index Builder (and an internationalization facility) for constructing indexes and a database search engine based on the company's patented PAT string search technology. The search technology allows users to locate, access, and display full text, search individual words and phrases located in multiple databases and file formats. The data within a single data base may be heterogeneous, with documents existing in different formats. TextSearch also includes the Parallel Execution Monitor (PEM).

PEM permits searching of different databases in parallel. Once a query is received, the PEM broadcasts the query to TextSearch engines, each running on a separate database. The results from the various TextSearch processes are collated and presented to the user.

The API for PEM and TextSearch are identical.

Open TextQuery is a multi-platform GUI that allows different levels of access or retrieval capabilities depending on user skills. Open Text Latitude is the most recent offering from Open Text. It is a client server system that uses Open TextSearch and retrieval technology to allow

users to search concurrently across multiple servers and databases without requiring the organization to structure and transform its data first.

Open Text supports SGML so that SGML data need not be converted to another format to be indexed and searched. What is required is the SGML text and the corresponding Document Type Definition (DTD), which contains structural information about the text. Open Text uses SGML structures during indexing and searching. (SGML and non-SGML data can be searched simultaneously.)

Open Text's engine supports major languages including those requiring multi-byte characters such as Chinese, Japanese, and Korean. Languages like Japanese with mixed characters (Kanji, Kana, and Roman) can also be searched. The search engine is the same binary version for all languages.

Platforms

The TextSearch server module runs on DEC MIPS (Ultrix), DEC Alpha (OSF/1), Hewlett Packard 9000 (UX), IBM RS6000 (AIX), Microsoft Windows NT, SGI, and Sun SPARC (Solaris 2.x, Solaris 80x86, and SunOS 4.1.x).

The client interface TextQuery is available for ASCII terminals, Macintosh, MS Windows, and OSF Motif.

Technical Fact Sheet

Indexing

Indexing in TextSearch is based on Open Text's string processing technology (called PAT technology) of indexing/searching strings of characters rather than words. Indexing is based on Patricia trees or suffix arrays in which documents are viewed as one long string where each position may be the start of a semi-infinite string which is defined by a starting position and extends to the right through the end of the text. (Two semi-infinite strings are not equal if their starting positions are different.)

A problem with using suffix arrays is one of determining the starting position for indexing, i.e., the position of indexing points. Indexing points can be placed at the beginning of any word or at every character. For languages like English, French, Italian, and Spanish where there is relatively little compounding, index points should be positioned at the start of a word which is the default for Index Builder. For languages where compounding is common (e.g., German) or large alphabets (or sets of characters) and no white space between words, the determination of index points is difficult. Although Open Text is developing index-point locators for a number of languages, its internationalization module currently supports German and Chinese. Moreover, Index Builder does have a facility allowing for database-specific identification of index points by Open Text's users. Index Builder supports lists of stop words that are customizable by users. However, Open Text does not recommend the use of stop words, since they can provide contextual information. Moreover, based on the string indexing and retrieval methods employed

by Open Text, there is no performance degradation when searching for phrases containing stop words as compared with queries with less common terms.

TextSearch Index Builder can index several megabytes of text per minute for collections that can fit into the computer's main memory.

Typically, index size is 50 to 75 percent of the original collection. (Text can also be compressed to about 25% of its original size with no effects on searching, since searching is not performed on text.) Indexes are automatically compressed with no appreciable degradation in retrieval speed due to the trade-off between the amount of compressed data that can be transferred per block disk read and the performance of the decompression algorithm.

Searching

Searching capabilities in TextSearch include:

Keyword

TextSearch supports keyword searches including case-sensitive and case-insensitive queries, as well as verbatim searches of phrases that also contain stop words.

Boolean

The Boolean operators AND, OR, NOT, and parenthetically-grouped combinations of Boolean operators can be used in queries.

Proximity

Adjacency searching is built into the key word search capability.

Bi-directional adjacency is also supported.

Directional proximity (where one term must follow another in a search) and bi-directional proximity (where order is irrelevant) are used to find terms within a specified distance of each other.

Searches can be restricted to parts of documents such as sentences or paragraphs.

The TextSearch API provides a command for determining frequently-occurring words and phrases in a database. This allows users to find the words or phrases that follow a phrase such as "the association of."

Thesaurus Support

Although Open Text does not provide a thesaurus, users can include their own or a third party thesaurus for searching.

Acronym Searches

Through TextSearch's thesaurus support, acronyms can be used in searches.

Misspellings

Users can modify TextSearch to allow for searching text for misspelling and typographical

errors. A searcher specifies a mismatch as either an insertion of an extra character, a deletion of a character, a replacement of a character, or a transposition of characters in a word.

Weights

Terms in searches can be weighted, and results are ranked based on highest or lowest scores. TextSearch supports searches of structured and unstructured document components in the same query and each part of the query, can be weighted differently.

Documents can be ranked by dates by assigning weights to most recent documents.

Date Searches

Documents can be retrieved for a range of dates, before a certain date, or after a certain date.

Restricted Searches

Searches can be restricted to the most recent set of retrieved documents.

Document Counts

TextSearch can be configured to return the number of hits rather than the documents themselves.

Distinguishing Features

The central feature of Open Text is its string search technology (rather than an inverted index approach) that allows for high performance indexing and searching. Open Text provides international language support including those languages based on multi-byte characters such as Chinese, Japanese, and Korean. Open Text can process SGML text without the need for conversion. SGML structures are used both in indexing and searching.

Personal Library Software

Headquarters:

Personal Library Software, Inc.
2400 Research Blvd., Suite 350
Rockville, Maryland 20850-3243

Tel: 301.990.1155

Fax: 301.963.9738

email: info@pls.com

WWW: <http://www.pls.com>

Corporate Fact Sheet

Business

Personal Library Software (PLS), Inc., develops and markets computer software for managing data, text, and image bases. PLS's products are being used for on-line databases, CD-ROM databases, corporate databases stored on magnetic and optical media, and for databases created by scanning and optical character recognition.

Components of PLS's products grew out of work on a system called SIRE that was developed by Mathew Koll and his colleagues at Syracuse University.

PLS was founded in 1983.

Markets

PLS products are used in many sectors including computer manufacturing, oil companies, libraries, pharmaceutical companies, newspaper and magazine publishing, law firms, military and intelligence agencies, defense firms, engineering companies, and utilities companies.

Some Key Alliances and Customers

The PLS search engine is embedded in the products or services for a wide range of companies including:

America Online (AOL)—PLS provides the core search technology for AOL; AOL remarkets PLS's text retrieval software as part of AOL's technology licensing program.

Apple Computer—PLS's engine is the core of the AppleSearch product.

Appleton & Lange (which acquired Macmillan New Media)—Macmillan New Media's entire medical product line on CD-ROM, which includes the titles AIDS Compact Library, The New England Journal of Medicine, and The Physician's MEDLINE, uses PLS's retrieval software.

Grolier Electronic Publishing, Inc.—PLS provides the search engine for the 1995 Grolier Multimedia Encyclopedia and the 1993 Guinness Multimedia of Records.

DataTimes—PLS provides the core search technology and user interface for electronic library systems for newspapers.

Knight-Ridder Information (Dialog)—PLS's products are used in WWW applications for customer service and support; Knight Ridder Information has a minority investment in PLS and is represented on the PLS Board of Directors.

Other key relationships that have been established include Congressional Quarterly, Dow Jones, Financial Times Information Services, NewsNet, Prodigy, and Time Inc.

The U.S. House of Representatives' information service on the WWW (<http://www.house.gov>) also uses PLS's search software.

Products

Core Technology

Personal Librarian

The core of PLS's products is its Personal Librarian (PL), which is a suite of electronic publishing, authoring, and retrieval tools. It consists of a full text search engine and a graphical user interface.

Maximum number of records: 16 million

Maximum fields per record: 255

Maximum record size: 2Gb, databases can be compressed

Maximum database size: main index limited to 2Gb

Index overhead: ~35%-45%

System Hardware Requirements

IBM PC 386 or higher (at least 2MB of memory, hard disk, and DOS 3.3 or later)

Sun-4[™]

VAX

Mac II or higher

Related Products

Callable Personal Librarian (CPL)

Applications can be developed using Callable Personal Librarian (CPL), a C language API. It

includes both the document-indexing functions and the search and retrieval functions. CPL is written in C and C++ and is designed to run under UNIX, VMS (5.3 and later), IBM-compatible microcomputers under DOS (3.3 and later) and MS Windows 3.x as a DLL), and on Apple Macintosh. CPL has no special library requirements other than the host system's C run-time library. It can also be used for creating text-based and graphical interfaces.

PLWeb

PLWeb, which uses any WWW browser, provides the tools to set up a server, create databases, and access information. It supports distributed searching, login and password control, per-document pricing, and billing and tracking logs. PLWeb requires WWW browser clients with forms capabilities.

Access to a demonstration of WebServer can be found on the homepage for PLS, <http://www.pls.com>.

Performance

On a SunSPARC 10 or 20, CPL indexes about 150MB of text per hour. On a 33MHz 386 PC with extra memory and disk caching, indexing is typically at 8MB per hour.

Technical Fact Sheet

Search Techniques and Query Language

PL offers both Boolean and natural language queries. For natural language inputs, PL eliminates stop words from the string and ORs together the remaining words. Concise queries loaded with relevant words work best.

PL provides relevance ranking capabilities using word counts in each document, word frequency in a document, document length, and number of term frequencies in the entire database compared to term occurrences in each document.

Users can also use relevant documents for a subsequent query. PL compares word occurrence in the relevant document with similar word frequencies in other documents.

A statistical query or concept search is one where PL first generates a list of terms that are statistically related to the words in a query. Those words that have a significant degree of co-occurrence with the query are deemed related within the context of the database. The concept search then performs a conventional search using the original query words and the 20 most significant related terms.

The PL language supports:

Standard Boolean operators—The standard Boolean operators AND, OR, and NOT are available.

Relevance ranking—Documents retrieved by a query are ordered by computed probability of relevance to the query.

Proximity/adjacency—Both proximity (within N words) and adjacency are available.

Statistical thesaurus—CPL has a set of tools that suggest possible new words based on words or documents the user already knows. This is based on co-occurrence relationships of terms in the documents in the database. See expand search below. (There is no preprocessing or specialized indexing required to use this feature.)

Document as query—CPL has a query operator that locates documents similar to a given one.

Field searches—Documents can be structured into fields, and searches can be restricted to certain fields.

Field display—Display modes can be set to named fields.

Wild card search terms - Word variants can be searched using wild cards.

Word stemming—CPL will automatically include variations of single or multiple word bases in the query. Stemming is restricted to singular, plural, and verb endings.

Synonym thesaurus—CPL supports the use of a user/administrator defined synonym thesaurus. A query operator controls the replacement of a word with a synonym.

Past queries—Past queries can be saved and reused to form new queries or new combinations of queries.

Range search—Range searching operations are allowed for numerical information. This is used for greater than, less than, and equality in fields containing specified values.

Expand search—PL can suggest search terms to the user. It produces a list of words that are related to the words in the query based on co-occurrence with terms in the database.

Differentiating Features

PLS provides 16-bit character encodings which lends itself to multilingual capabilities. A statistical thesaurus that provides alternate search terms based on co-occurrence relations helps users in their searches. PLS also supports searches on databases while they are being updated.

Miscellaneous Information

Foreign language support—CPL is currently being used in French, German, Italian, and Spanish, language applications.

Quality Information Systems, Inc.

Headquarters

Quality Information Systems, Inc.
10680 West Pico Boulevard, Suite 260
Los Angeles, CA 90064

Tel: 310.287.0728 or 800.95-Doc-Mgmt
Fax: 310.287.1622
email: QIS_Fleiss@earthlink.net
WWW: Under construction

Corporate Fact Sheet

Business

Quality Information Systems (QIS) develops and markets search technology and document management systems.

QIS was founded in May 1992.

Markets

QIS's initial market has not been restricted to any particular segment.

Products

Core Technology

The core component of QIS products is the Full Access Search Tool Engine (F.A.S.T.), which is a set of C routines callable by either a C or C++ application. F.A.S.T. also includes a 250,000 word dictionary which is used in indexing and searching as described below. The dictionary includes antonyms, homonyms, and synonyms for entries, as well as stop words. (Stop words comprise about 2% of the entries.)

The dictionary was developed by QIS.

The F.A.S.T. Engine provides a set of routines for dictionary management (stopword designation, control of user and predefined dictionaries, compress/decompress, and synonym

definition), document management, optical character recognition support, document marking, and search.

Related Products

QIS expects to release its document management system called Full OnLine Documentation Information Retrieval (FOLDIR) toward the end of 1995. It will include the F.A.S.T. search engine.

FOLDIR will provide functionality for obtaining information on changes to the document collection (audit trails), document flow (creation, review, modification, and editing of documents), change control (check-in/check-out data), and version control.

FOLDIR will be Internet compatible where documents are translated from their native format (e.g., Word, WordPerfect, Lotus Notes, SGML, or Excel) to HTML+ during check-in. (On check-out, the documents are translated back to their original formats or to HTML.)

A scripting language will also be provided.

FOLDIR will support over one trillion document components, and each document can contain over 16 million bytes.

Platforms

Currently, F.A.S.T runs on Microsoft Windows 3.1, Macintosh, and Windows NT. It will be ported to UNIX platforms later in 1995. (F.A.S.T. is implemented in C with ANSI features only.)

Technical Fact Sheet

Indexing

For indexing, FAST converts documents to an HTML+ format and builds an inverted index for all non-stop words in the document. (The choice of HTML+ format was a design decision to simplify the tasks of building a viewer, marking documents for specific applications, and supporting aspects of proximity searching.) Dictionaries (predefined or user defined) are used in the indexing to locate occurrences of non-stopwords in all the documents. Links are made between the words in the inverted index and the entries in the dictionary during indexing.

All dictionaries are partitioned into fourteen sub-dictionaries based on the numbers of letters in the words. Words with 14 characters or more are stored in the same dictionary. The dictionaries are sorted alphabetically and are searched using a binary search algorithm.

By default, characters are upper and lower case alphabetic characters, but users can include numeric and four special characters for indexing purposes.

Searching

Searching capabilities include:

Word searches—Query terms must match exactly.

Fuzzy word searches—Searches for words with some variation. A threshold parameter for constraining or broadening the degree of fuzziness of searching can be modified.

Prefix searches—Searches for words that begin with two or more characters is supported.

Synonym searches—Based on the predefined or user dictionaries, searches can include synonyms of search terms.

Numeric searches—Numeric information and relations (e.g., greater and less than and equality) are supported. Users are also able to search for all numbers that are within a range of certain specified numbers.

Proximity searches—Searches for pairs of terms within a specified distance and order are supported. Searches can also be restricted to sentences and paragraphs.

Text string searches—Predefined text strings can be used in searches.

re:Search International

Headquarters

re:Search International
One Broadway
Cambridge, Massachusetts 02142

Tel: 617.577.1574
Fax: 617.577.9517
email: rintl@world.std.com
WWW: <http://www.research.com>

Corporate Fact Sheet

Business

re:Search International develops and markets authoring and retrieval software. It is a division of the consulting firm Toronto Corporation.

The company was founded in 1983 and was called Retech. In about 1990, Retech was purchased by a group of investors and became MicroRetrieval Corp, and in 1994 became known as re:Search International.

Markets

re:Search has targeted the CD-ROM publishing community as its major market.

Sample Applications

re:Search is used in the following CD-ROM titles:

Wm. C. Brown Publishers—*Life*
USDA Farmers—*FMHA Instructions & Forms*
IEEE—*IEEE Conference on Neural Nets*
IEEE—*IEEE Conference on Acoustics, Speech, & Signal Processing*
U. S. Joint Chiefs of Staff—*Joint Warfighting Publications of the Armed Forces*

Products

Core Technology

re:Search—re:Search, the core technology of re:Search International, consists of an indexing module and a search module. The indexing module allows users to categorize or classify documents according to subject area and creates one or more indexes for subsequent searching. The search module permits users to browse documents or conduct a search of a collection of documents.

Image retrieval, which is included in re:Search, allows the user to browse and search for images including photographs, charts, and drawings, from a document in context with the text.

Related Products

API Toolkits—There are toolkits for Windows, DOS, and Macintosh allowing re:Search's engine to be embedded in user applications.

Platforms

PCs—286/33Mz (512k, 640 preferred) under DOS 2.72
386/25Mz (512k, 640 preferred) under Windows 1.43

Macintosh—Mac IIcx or higher (4MB) under MacOS 7.x

Sun—SPARC (8MB) under Solaris 2.x, SunOS 5.2

Specifications

Maximum catalogs and document pages: Limited only by available storage

Maximum documents in one catalog: Limited only by available storage

Maximum number of fields per catalog: 255 (100 for DOS)

Technical Fact Sheet

Indexing

re:Search permits full text and field indexing using inverted files. A stop word list is provided and can be edited.

re:Search can index up to 40MB per hour.

Searching

Searching returns a ranked list of documents with highlighted hits that can be browsed. The

ranking is determined by the number of hits within a document.

Single word or phrases—Users can enter a single word or a phrase as a query. Numeric information can be included in a search.

Wild card expansions—Users can truncate terms in a query with wild cards for searching segments of words. Prefix, suffix, and embedded wild cards are allowed. Wild cards can be used with other types of searching including Boolean.

Support for Boolean operators—Users can use the Boolean operators AND, OR, and NOT to constrain searches.

Proximity word searches—Both uni- and bi-directional proximity searches are supported. Bi-directional search is the default. re:Search permits multiple proximity searching where users can specify the number of words between terms in a query.

Fielded data support—Users can search for particular parts of documents such as section titles by using field search. Up to 255 fields per catalog (fielded and full) are permitted.

Search locally within a document—Searching can be limited to a single document rather than a collection of documents.

Toggle between rank and document alphabetic ordering of hits—Users can view the set of returned documents either alphabetically or by ranking.

Save searches—Searches can be saved and re-used in later queries.

Hypertext links (Windows version)—Hypertext links connect current search or browse text to related text either in the same document or in another document in the collection.

TextWare Corporation

Headquarters

TextWare Corporation
P.O. Box 3267
Park City, UT 84060

Tel: 801.645.9600

Fax: 801.645.9610

email: Not available

WWW: <http://www.textware.com>

Corporate Fact Sheet

Business

TextWare, Corporation develops and markets full-text indexing and retrieval software. It also offers authoring and indexing CD-ROM production services.

Textware was founded in 1989.

Markets

TextWare directs its marketing to those implementing on-line databases and electronic publishing applications, especially CD-ROM authoring. TextWare markets a product called TextWare Lite which is a royalty-free module that can be distributed freely in applications built by TextWare's customers.

Sample Application

Federal Deposit Insurance Corporation (FDIC)—The FDIC, including bank examiners, uses TextWare's technology for accessing its Rules & Regulations (which originally consisted of over 10,000 pages of detailed policies) while doing field audits. Later, FDIC's Division of Supervision added the Manual of Examination Policies, Regional Director Memoranda, and other data as TextWare databases and the Division of Liquidation has implemented a similar database.

Products

Core Technology

TextWare is the indexing and search engine enabling users to create, organize, and search text

databases which are called Cardfiles. CardFiles consist of one or more text files (called Cards) that are indexed by TextWare.

Related Products

TextWare offers three other products, TextWare, TextWare Retrieve Only, and TextWare Lite. TextWare Retrieve Only is a retrieval component that is used to distribute Cardfiles to users who do not require indexing capabilities. TextWare Lite is a subset of TextWare Retrieve Only that is used for electronic publishing and comes with a royalty-free licensing arrangement.

System Requirements

TextWare runs under DOS, Windows and OS/2 and requires 512KB of RAM and 2.5MB on the hard disk.

TextWare Lite runs under DOS, Windows, and Macintosh and requires 512KB of RAM and 1.2MB on the hard disk.

Technical Fact Sheet

Cardfiles

TextWare allows users to organize large unstructured documents or collections of documents into databases called CardFiles. A CardFile can be a single text file, a part of a file, e.g., a paragraph or a few lines of a file, or a collection of files. In turn, CardFiles are automatically divided into smaller parts, e.g., a file, a page, or a paragraph, and indexed by TextWare.

Users can link Cards together to form a Group which represents related information. A Group, for instance, can be sequential sections of a manual.

TextWare allows users to search across multiple CardFiles, all Groups within a CardFile, or one or several Groups.

Indexing

Indexing, based on a TextWare proprietary algorithm, proceeds when a Card File is defined. Each word is indexed and, because indexing is at the level of CardFiles, new files can be added without re-indexing an entire CardFile.

While Word for DOS, WordPerfect, and ASCII files can be left in their existing formats for indexing, TextWare converts other file formats into its ASCII-like format in producing a CardFile. (Over 40 file formats are accepted.) CardFiles can also contain database records.

Foreign characters that are part of the extended ASCII character set and non-alpha-numeric ASCII characters (e.g., dollar signs) can be included or excluded in the indexing process.

Synonyms and stop lists can also be defined.

A CardFile can be up to 2GB of text with no limit on the number of files and words in a CardFile.

Indexing Time and Size

CardFiles can generally be indexed at up to 40MB of text per hour on an 80-386 PC with 4MB of extended RAM. TextWare can use up to 32MB of extended RAM for indexing purposes.

Indexes are typically 10 to 20 percent of the size of the original text. When the text files have been converted to the ASCII-like internal format, both the CardFile and its index can be compressed to approximately 65 percent of the size of the original data.

Searching

Searching in TextWare can be done across 64 CardFiles simultaneously, or restricted to specific CardFiles.

A query can consist of up to 200 characters, and wild cards can be used for one or more characters.

Boolean searches (AND, OR, NOT, XOR, ANDNOT, and ORNOT) are supported, as is uni- and bi-directional proximity searching. The operator XOR represents exclusive OR, where the query A XOR B succeeds if a document (a card) contains the term A or the term B, but not both. The query A ANDNOT B matches all documents containing the term A and not the term B. The query A ORNOT B finds documents that contain A and all documents that do not contain B.

Searches can be saved and re-used or appended to in subsequent searches. Term hits or their synonyms are highlighted in the retrieved list.

TextWare also provides users with a wordwheel that displays a CardFile's index so that users can select appropriate query terms.

TextWare does not perform any morphological analysis or stemming.

TextWare does not include a thesaurus or an interface to one.

TextWare has no foreign language capabilities.

Thunderstone Software

Expansion Programs International, Inc.

Headquarters

Thunderstone Software
Expansion Programs International, Inc.
11115 Edgewater Drive
Cleveland, Ohio 44102

Tel: 216.631.8544
Fax: 216.281.0828
email: info@thunderstone.com
WWW: <http://www.thunderstone.com>

Corporate Fact Sheet

Business

Thunderstone Software develops and markets text retrieval software. The company was founded in 1982.

Markets

Thunderstone markets its technology to developers of large organizations that are building information management and retrieval applications. These include government agencies, financial institutions, consulting firms, and law offices.

Sample Applications

Applications that use Thunderstone's software includes:

Dow-Jones: Telerate News Agent, a news retrieval system with 80,000 users on Wall Street.

Dow-Jones: NT News Agent, a system for profiling, disseminating, and retrieving information via Lotus Notes.

Novell, Inc.: WWW Server—on-line searches for information concerning Novell and its products and services. See <http://www.novell.com> or <http://netware.com>.

Starwave (Paul Allen): WWW Server—on-line access to *Outside Magazine* and ESPN.

WordPerfect—Softsolutions: Intelligent Search, a package for information retrieval within the Softsolutions Document Management System.

Thunderstone has also provided technology for applications involving Dow-Vision, Desktop Data, Arthur Andersen & Co., Chemical Bank of New York, and the NASA Johnson Space Center.

Products

Core Technology

METAMORPH

METAMORPH is a collection of real-time pattern-matching algorithms used in searching. METAMORPH has an editable thesaurus of over 250,000 word associations that is used for query expansion. METAMORPH also contains a morphological analyzer to determine roots of terms in a query.

Related Products

TEXIS

TEXIS is an SQL-compliant relational database server for managing textual information. There is a zero latency insert in which newly added documents are immediately searchable. These items are searched linearly until the index is updated. There are variable sized records and variable length fields. Although it is not recommended, a field can contain up to 1 Gigabyte of information. TEXIS also uses variable length index keys based on Thunderstone's variable length key Btree.

3DB

3DB is an indexed text retrieval system that is optimized for search speed and allows for efficient searching of collections that exceed 10-20 megabytes. Many of METAMORPH's features are included in 3DB except that 3DB does not contain a thesaurus. However, subsets of text in a 3DB database can be passed off to METAMORPH for refined searching using a thesaurus. Each 3DB index can reference as many as 4 billion files, each containing up to 4 Gigabytes. Its indexes are typically lower than 15% of the original source.

APIs

The APIs allows C or C++ programmers to integrate Thunderstone's technology into customized applications. There are five such APIs, namely, the METAMORPH API, the 3DB API, the Network API, the Browser API, and the Network Code Generator.

The METAMORPH API enables developers to integrate METAMORPH indexed concept-based text searches with applications. This is particularly useful for dynamic searching and message handling.

The 3DB API is used for developing indexed databases for large text collections. It is also used for retrospective searches.

The Network API is the client-server version of both the METAMORPH and the 3DB APIs.

The Browser API allows for integration Thunderstone's file browser into DOS, UNIX, Windows, and X-Windows applications.

TEXIS Mosaic SQL Bridge

TEXIS Mosaic SQL Bridge translates HTML formatted queries into an SQL query processed by TEXIS. It then converts data retrieved by TEXIS into HTML format which is sent to the Mosaic server.

Other Thunderstone products include the Network Code Generator (NCG) for supporting the development of client/server programs, ITS-WRITER for generating interactive scripts and menus, FILE-FINDER which is similar to UNIX's "find" command, the Postscript viewer which incorporates METAMORPH with Ghostscript, and METABOOK which supports retrieval applications involving book or manual information. (METABOOK generates a table of contents automatically.)

Platforms

Thunderstone products run on DG-MV, Macintosh, MS-DOS, Windows, and Windows NT, OS/2, MVS-XA, and UNIX.

Benchmarks

METAMORPH can search up to 4.5MB per second for linear search. It takes up to 4.5MB a minute to create an index.

TEXIS can have up to 10,000 tables per database with 4 Gigabytes as maximum table size. Chained tables are permitted, thereby removing the 4 Gigabyte limitation. The limit on rows per table is 1 billion and there can be 50 columns per table as well as 50 indexes per table. Maximum field size is 32 characters and the maximum tables per query is 32. Maximum index key size is 8192.

Indexing overhead is about 15% of original text.

TEXIS approximate indexing rates

Intel 486-33Mhz	1.4MB/min
SPARC II	1.9MB/min
SGI 4D/35	2.5MB/min
HP800(1 processor)	3.6MB/min

Technical Fact Sheet

Indexing

Texis uses an inverted index method and is capable of handling structured data attributes. Non-

text objects can be indexed through a textual description of the object. The contents of an index, i.e., a wordlist can be viewed by a user.

A stop list of about 150 words is provided and can be modified by the user.

Morpheme processing is performed at query time and not at indexing time.

Searching

Thesaurus

Thunderstone views searching using Metamorph as one of set logic. Words in a query are expanded based on items in the thesaurus which is included in the product. For instance, in a query using the terms "Near East" and "power struggle" both terms would be expanded in a set of associations. "Near East" has six associations including Kuwait. "Power" is expanded into 57 possible terms, including "force," whereas "struggle" has 23 associations including "conflict." Metamorph then searches for documents that contain at least one term from each of these association lists, and returns those documents containing a term from each set. (In Thunderstone's terminology, Metamorph is searching for an intersection of the three lists. Set logic is described below.)

The built-in thesaurus, constructed by Thunderstone, has over 250,000 entries. The thesaurus can be customized by users.

The thesaurus contains parts of speech for each of its items and the categories are taken from the set: pronoun, conjunction, interjection, modifier, noun, preposition, verb, and unclassified.

When a word is stored as an equivalence to another root entry, Metamorph will back search for an associated root. For instance, suppose there is an item "A&R" which has an equivalence list "A&R, automation, robotics, automation and robotics." A search for "robotics" will automatically associate "A&R" as an associated term for "robotics." This feature is called "back-referencing."

Metamorph also supports "see referencing" where one term may point to another term in a different association list. For example, consider the association lists "cowboy, horse, cows, rancher" and "rancher, plains, landowner." A search for "cowboy" expands to "cowboy," "horse," "cows," and "rancher." With see referencing invoked, "cowboy" not only expands to "cowboy," "horse," "cows," and "rancher," but also to "rancher," "plains," and "landowner." To prevent looping and excessive hits, see referencing can only be applied to one level and sets of equivalences are truncated at 256 terms.

Based on multiple word entries in the thesaurus, idioms and compounds can be identified in text by the longest consecutive sequence of terms matching an entry in a thesaurus.

Users have the option of using term expansion in their searches.

Set Logic

As noted above, searching in Metamorph is performed in terms of set logic involving an intersection of the thesaurus-based lists of the terms in the query. So a query A AND B AND C, where A, B, and C are terms and AND represents intersection is the default mode of searching in Metamorph. This example is a query with three terms and two intersections, and all three terms must be satisfied if the query is to succeed.

In the same query A AND B AND C, a user could "loosen" the search to a single intersection rather than to two. A parameter could be sent and Metamorph would interpret the revised query as either A AND B, A AND C, or B AND C. Thus, if any one intersection holds, the query will return the matching documents.

Users can use the same logic to search for no intersections, which behaves like an exclusive OR.

While each term in an intersection is treated as the same in that each term in an intersection must be present for the query to return a hit, a user can specify a query to ensure the presence of one term and at least one of the other terms. That is, not all terms must be present. This has the same effect as a Boolean combination of an AND and ORs.

Negation is also supported by Metamorph where a term in a query or its associations from the thesaurus are marked for exclusion. That is, the marked term must not appear in any of the retrieved documents.

The logic operators can be used in combination with each other.

Other Features

Proximity ranges can be adjusted by users. In the default case, sentences are returned as hits during searching. However, searchers can adjust proximity ranges to within a sentence, a line, a paragraph, or a document. Searches can also be restricted to character ranges.

A morphology engine determines roots of inflected words by removing prefixes and suffixes of words. Keyword search is done on roots of words. (When a search term consists of two or more words, morpheme processing applies only to the last word in the phrase.)

Although stopwords are generally removed from a query, Metamorph supports phrase searching (e.g., searching for "state of Massachusetts") where order is preserved and stopwords are included in the search process.

Wildcard searches of up to 80 characters can be used. Multiple wildcards within a single search can also be used. Wildcards can be used inside a word or for finding one term near another term.

A numeric pattern matcher locates numeric information and English expressions of numeric quantities. Equivalent numeric expressions can be found. For instance, METAMORPH can determine that "thirty-four dollars and 15 cents" is the same as "\$34.15" and that "four score and seven" is 87.

An approximate pattern matcher locates deviations, typographical errors, transpositions, and misspellings. The user can select the percentage of term deviation that is tolerated in a search. The default is to locate terms that are at least an 80% match. The approximate pattern matcher is case insensitive, works with alphanumeric characters, and can be used with logic operators.

A regular expression matcher locates any fixed or variable length expression such as social security number, dates, chemical formulae, footnotes, or special headers.

Metamorph also supports relevance ranking and term highlighting.

Differentiating Features

Thunderstone offers different products for searching document collections of various sizes. Taxis is used for collections under 20 MB, while 3DB is used for databases exceeding 10 - 20 MB since it is optimized for search speed. Thunderstone also offers query expansion based on their relatively large thesaurus.

Miscellaneous

Expansions Programs International, Inc., is privately held and was incorporated in 1981 in California. The company was originally involved in human factors concerning information processing in the Pacific Rim area. In 1984, the company moved its headquarters to Ohio and began work on text processing applications under the trade name of Thunderstone Software.

Thunderstone is a small business and is registered with the U.S. Small Business Administration. All the original developers are currently with the company.

TMS, Inc.

Headquarters

TMS, Inc.
206 West Sixth Avenue
P.O. Box 1358
Stillwater, OK 74076

Tel: 405.377.0880
Fax: 405.377.0452
email: info@tms-ok.mhs.compuserve.com
Web: <http://www.tmsinc.com/tms>

Corporate Fact Sheet

Business

TMS, Inc. develops and markets text and image technology for use in the development of text and image retrieval, workflow, CD-ROM publishing, engineering drawing systems, and fax server products.

TMS was founded in 1981 by Dr. J. R. Phillips, former chairman of the Computer Sciences Department at Oklahoma State University. It was the first company to demonstrate full-text and hypertext retrieval designed specifically for navigating databases on optical media.

Markets

TMS markets to large corporations, system integrators, and VARs. TMS specifically targets organizations that publish CD-ROMs of technical and engineering manuals and catalogs or have retrieval applications with large imaging requirements.

Sample Applications

Price Waterhouse—Field auditors for Price Waterhouse use TMS products for searching for information from reference materials while performing audits.

Other customers using TMS's retrieval software include DiscoverCard, General Dynamics, Rockwell International, University Microfilms International, and U.S. Navy.

Products

MasterView is an API for providing text and image retrieval capabilities including document imaging and hyperlinking (text-to-text, text-to-image, image-to-text, and image-to-image). Links to relational databases, multimedia, OCR, bar coding, and other applications can be established using MasterView.

TMS Database Preparation is the indexing component for MasterView.

ViewDirector is an imaging toolkit providing an image manager/viewer and supports the addition of black and white or color image capabilities to new or existing applications. It features software compression, decompression, conversion, and manipulation for a variety of industry standard file formats. A scanner API is also included for creating bitonal scanning applications.

Platforms

MasterView and ViewDirector are available for Microsoft DOS, Windows, and Windows NT, Apple Macintosh (Mac OS 4.5 or higher), OS/2, Sun SPARCStation (SunOS4.1.x or Solaris 2.x), IBM RS/6000 (AIX3.2 or above), and HP9000 Series 700 and 900 (HP-UX 8.0.7 or higher).

Technical Fact Sheet

Indexing

Before indexing, documents are converted to a TMS-proprietary markup language, thereby making the structure, attributes, and hierarchical nature of the documents explicit. Indexing is performed on the annotated documents using an inverted file as the index.

Indexes are optimized for speed and storage efficiency. A typical index is 30%-60% overhead.

Indexing rates on full text collections can exceed 50MB per hour on a PC 486/66.

Searching

Search capabilities include:

Single word or phrase searching—Single words or phrases can be used in a search.

Boolean operations—Any combination of the operators AND, OR, and NOT can be used in a query.

Right truncation—Users can truncate the right part of a word in a search.

Field search—Field searches of up to 256 fields can be used.

Restricted search—Users can restrict searches to the previous search list hits rather than the entire collection of documents.

Proximity searching—Searches can be performed either on terms within N number of words or for words preceded by N number of words.

Domain searching—Searching can be constrained to selected portions of a database. For instance, in a database containing a parts catalog and maintenance procedures, users can restrict searches to one portion of the database say, parts catalog, and not the maintenance procedures.

Ranking—Documents returned from a search can be ranked by the number of occurrences of a search term or by table of contents order.

Other searching functionality such as saved queries can be introduced through the API MasterView.

Verity, Inc.

Headquarters:

Verity, Inc.
1550 Plymouth Street
Mountain View, California 94043

Tel: 415.960.7600
Fax: 415.960.7698
email: info@verity.com
WWW: <http://www.verity.com>

Corporate Fact Sheet

Business

Verity, Inc. develops and markets technologies for filtering, searching, retrieving, and navigating information sources. The company's information technology, toolkits, and services are currently used in more than 650 corporations and organizations worldwide in a wide range of industries.

The company was founded in 1988.

Markets

Verity markets its technology specifically to:

Workgroups, departments, and corporations that need to organize and disseminate information.

End users who need to locate current or historical information.

Information and content providers, including on-line services, CD-ROM publishers, and news and data services, that need to make their information accessible to corporations and individuals.

Some Key Alliances and Customers

Verity products are used by:

Borland International Inc.—Verity's systems are used for customer support and problem resolution.

Hewlett-Packard (HP)—Verity's technology is at the heart of a system that matches resumes to job openings. HP's Corporate Communications Department is using Verity to develop a repository of information including press releases generated from each of HP's business units. Customer service for HP is also using Verity's products for their field consultants.

Lotus Development, Corp.—Verity's technology is the searching component for Lotus Notes.

Parke-Davis (prescription drug divisions of the Warner-Lambert Co.)—Verity's engine is used for searching new-product information in response to requests from regulatory agencies. The system is also used for accessing information on product labels.

Adobe Systems, Inc., and Frame Technology, Inc. have also embedded Verity's search engine in their products.

Products

Core Technology

Topic Development Kit (TDK)

Topic Development Kit (TDK) is the core of Verity's product line. TDK enables customers to embed Verity's search and retrieval capabilities within their products and to customize the user interface using the packaged API. This API is comprised of about fifty C language function calls, while TDK consists of a C-callable library, documentation, and sample applications.

Related Products

Topic Information Server

Topic Information Server allows agents to search for information located within an organization, as well as information found on the Internet. Server features include threads, symmetrical multi-processing, automatic load-balancing, and efficient network usage. There are also administrative capabilities such as automatic recovery, security, logging, compression, and encryption.

Topic Agents for Microsoft Windows, Motif, and Apple Macintosh

Topic agents are software objects that search, filter, and monitor information across an organization independent of document format, hardware and software platform, network, and location. Agents can perform retrospective searches or pro-active ones against multiple data sources, e.g., Lotus Notes databases, word processing documents, or newswires.

Topic Information Server for the World Wide Web

The World Wide Web server uses Verity's search and retrieval technology for navigating the Web.

Demonstrations are available at <http://www.verity.com>.

Topic Agents for Mosaic

Topic Agents for Mosaic allows Mosaic to use agents stored on local and remote sources, giving users easy access to distributed information.

Topic Agents for Lotus Notes

Topic Agents for Lotus Notes enables users to locate information residing in multiple Notes databases.

Topic Agents for Microsoft Exchange

Topic Agents for Microsoft Exchange allow users to search for information residing in Microsoft Exchange databases.

Topic News Server

The Topic News Server handles newswires to provide personalized information to its users. It filters information from news services and provides for dissemination of news and on-line subscription services. It replaces a previous product from Verity called Topic Real Time.

Topic Document Entry

Document Entry, which is available for Microsoft Windows and Motif, allows Topic users to add documents to Topic databases. Documents can be added while other users are accessing the database.

Strategic Alliances for the World Wide Web

Verity has established relationships with several third party software companies, e.g., Adobe, Interleaf, Spyglass, and Terisa Systems, to provide customers with internet publishing technology.

Supported Operating Systems

- UNIX (Sun, HP, OSF, IRIX)
- OS/2
- DOS
- Windows
- Macintosh

Supported Platforms

- DEC
- IBM
- Intel
- Sun Microsystems
- Hewlett Packard
- Apple Macintosh
- SGI
- Many others

Supported Network Vendors

- IBM
- DEC
- Sun Microsystems
- Apple Computer
- Novell
- 3Com

Technical Fact Sheet

Search Techniques and Query Language

Verity's tools provide several options for indexing documents, as well as for searching or navigating collections. Each will be described in turn.

Indexing Methods

A document for TOPIC may consist of text, either ASCII or a wide variety of word processing formats, fields of structured data, annotations, relationships to other documents, files or images, and/or TOPIC queries. These relationships are expressed as Hyperlinks. TOPIC products distinguish two parts of a document, namely, structured data, e.g., title and author, and full text. The structured data, which is extracted using a lexical analyzer, can be searched separately or in combination with full-text. Large documents can be decomposed using pattern matching or SGML tags, and indexed and search separately if desired.

The indexing format stores full word indexes and is proprietary to Verity. Optimized for speed, the format stores the size, location, and count for each word instance in a collection.

There are three options for indexing text. The first allows users to specify that non-alphanumeric characters (e.g., "&" and "-") should be included in the index, and therefore, be searchable. A query, for instance, for "AT&T" would be found if "&" is specified as a indexable character. TDK supports user-definable lists of indexable non-alphanumeric characters. By default, non-alphanumeric characters are not indexed.

Although a list of stop words are included in TDK, users can supplement the list by including other words that are not meant to be indexed. These entries could be words that occur with high frequency in a particular collection. Through the TDK, users have control over the terms in the exclusion list. These terms can be listed explicitly or specified using regular expressions. Words can also be excluded based on length.

Indexing can also be restricted to a narrow vocabulary with another option where users specify only the terms that they want to include in an index. The result is a small and very specific index. Through the TDK, users have control over the terms in the inclusion list. Terms can be listed or specified in terms of regular expressions.

TOPIC uses a character pattern analyzer and extensions for regular expressions to extract data values from documents. The type of data values that can be identified include words, sentence boundaries, newlines, paragraph boundaries, a sequence of one or more tabs, white spaces, and punctuation. For information residing in specific field, TDK has the capability of indexing complex types such as dates and numerical information or ranges.

Search Rules and Concept Operators

Searching in TOPIC can be done in several ways, including a keyword or concept-based approach. Concept-based searching involves topics which are hierarchical groupings of information of some domain or subject area. They consist of three components, namely, structure, weights, and operators. Subtopics are those levels below the root node called a topic and the lowest levels or leaves which are called evidence topics. Evidence topics are alphanumeric strings.

A structure can, for example, define the concept or topic "airline services" as being related to the subtopics "frequent flyer programs," "vacation packages," and "reduced rates."

Weights together with operators express the importance of each piece of evidence in a TOPIC tree and are used to calculate the scores for parent and a child topic during a search. In other words, they affect the scores of hits. A weight that is assigned (by a user) reflects the significance of that child relative to its siblings that share the same parent. Weights must be a number in the range of 0.01 to 1.00 with higher numbers indicating greater importance. (They can only be assigned to the concept operators AND, OR, and ACCRUE.)

Weights not assigned by a user are automatically assigned by the system based on the operator involved. Subtopics take on the weight of 1.00 when the AND or OR operator is used. A weight of 0.50 is assigned when the ACCRUE operator is used. (Assigned weights will be changed if the operators are modified in accordance with the AND/OR and ACCRUE distinction.)

A document is scored during a search process by a bottom-up method of scoring based on the weights and operators in topic trees and proceeds as follows. A search first starts by analyzing a document collection against evidence topics in a particular topic. If the evidence topic is present, then that topic is given a score of 1.00; otherwise, the score is 0.00. However, if the evidence topics are weighted, TDK multiplies the scores of the evidence topics by their weights, and then combines the resulting products as specified by the operator of their parents. In the case that the parent of the evidence topics is a subtopic (i.e., has a parent), its score is multiplied by its weight, and the product is combined with the products of its siblings in accordance with the operator assigned to the parent topic. The process is continued until the root topic is reached.

The concept operators ACCRUE, AND, and OR play a key role in the scoring process and can be described as follows:

The ACCRUE operator takes two or more search items and returns a ranked list of documents containing at least one of the search elements. The use of ACCRUE rather than other operators impacts scoring.

The AND operator selects documents containing all search terms. It is similar to the ALL operator except that AND returns a ranked list of matches and ALL does not.

The OR operator selects documents containing at least one of the search terms. It is similar to the ANY operator except that OR returns a ranked list of matches and ANY does not.

In connection with these operators document scoring is determined in accordance with the following rules:

If a topic uses the ACCRUE operator, TDK takes the maximum of the products of each child's weight and score, then adds a little to the score for each child present in the document.

If a topic uses the AND operator, TDK takes the minimum product of the products of each child's weight and score.

If a topic uses the OR operator, TDK takes the maximum product of the products of each child's weight and score.

Documents returned are ranked by score or relevance. Documents displayed can be restricted to those above a user-defined threshold.

An Example

As an example, consider a possible topic tree for "transportation." The concept "transportation" could be defined in terms of the subtopics "railway," "shipping," "automotive," and "air." Suppose the weights for each of these concepts has the value .50, except that "air" has a weight of .75. Since any of the subtopics add supporting evidence to the main topic "transportation," the ACCRUE operator is appropriate to use with "transportation." In turn, the topic "air" could be defined in terms of two subtrees, namely, "international airlines" and "U.S. airlines" each having weights of .50 and .80, respectively, for the example. Also, suppose that the ACCRUE operator is associated with "air." The offspring for "international airlines" could be "British Air," "Japan Air," "Lufthansa," and "Swiss Air" each of weight .50. Since the (sub)topic "international airlines" could be satisfied if any of "British Air," "Japan Air," "Lufthansa," and "Swiss Air" are found, the operator that is most suitable is an OR.

To see how evidence, weights, and scores interact, consider a document collection that has an article about Lufthansa. The evidence topic "Lufthansa" would be matched, and by the rule which assigns scores to evidence topics, the topic "Lufthansa" receives a score of 1.0 whereas the other topics, namely, "British Air," "Japan Air," and "Swiss Air" receive a score of 0.0.

Since the operator of the parent topic "international airlines" is an OR and since the score for "international airlines" is the maximum product of the products of each child's weight and score, the score for "international airlines" is the product of .50 and 1.0, i.e., .50. (Recall that the score of all other evidence topics is 0.0.) Assuming no evidence for "U.S. airlines," the next calculation is for applying the rule for the ACCRUE operator which is the operator for the parent topic of "international airlines." The score for ACCRUE is the maximum of the products

of each child's weight and score plus a little that is added to the score for each child present in the document. Now the score for "international airlines" is the product of .50 (its weight) and .50 (its score from the calculation above), i.e., .25. Since there is no evidence for the other offspring of "international airlines," "U.S. airlines" receives no extra score. Similarly, supposing the scores for each of "railway," "shipping," and "automotive" is 0.0, the score for the concept of "transportation" with the ACCRUE operator is the product of the weight (.75) and score (.25) for "air" which is .1875.

As the example shows, the scoring process is bottom up from the specific evidence topics to the most general topics or concepts.

Search Operators and Modifiers

Besides the concept operators of ACCRUE, AND, and OR, TDK uses other operators in its search and scoring process also.

These include the Boolean, evidence, proximity, and relational operators. Verity calls ANY and ALL Boolean operators.

The ANY operator takes two or more search terms and returns documents containing at least one of the search terms. Documents returned are scored 1.00 and therefore cannot be ranked by relevance. This contrast with using the concept operator OR which ranks hits.

The ALL operator takes two or more search terms and returns documents that include each of the search terms. As with the ANY operator, documents returned are scored 1.00 and therefore cannot be ranked by relevance. This contrasts with using the concept operator AND which ranks hits.

The evidence operators are SOUNDEX, STEM, THESARUS, WILDCARD, and WORD.

The SOUNDEX operator selects documents that contain terms that sound like the query term. A query with the term "ocean" will match documents with words such as "oceanic" and "oceanographic." The matching documents are not ranked unless the modifier MANY (described below) is used.

The STEM operator returns documents that match one or more variations of a search term. The term "product" would also match "products" and "productions." The matching documents are not ranked unless the modifier MANY (described below) is used.

The THESAURUS operator finds documents that contains word that are synonymous with the search term. The matching documents are not ranked unless the modifier MANY (described below) is used. A thesaurus is provided with the system.

The WILDCARD operator returns documents that include matches to a character string containing variables. Special characters are used to designate different variable positions including a range of characters as well as exclusion of characters. The matching documents are

not ranked unless the modifier MANY (described below) is used.

The WORD operator finds documents that include at least one instance of the word in a query. The matching documents are not ranked unless the modifier MANY (described below) is used.

The proximity operators are PARAGRAPH, PHRASE, and SENTENCE.

The PARAGRAPH operator takes two or more terms and returns documents with those terms contained in the same paragraph. Search elements can be specified in random or sequential order. The matching documents are not ranked unless the modifier MANY (described below) is used.

The PHRASE operator takes two or more terms (in a fixed order, i.e., a phrase) and finds documents with that phrase. The matching documents are not ranked unless the modifier MANY (described below) is used.

The SENTENCE operator takes two or more terms and returns documents with those terms contained in the same sentence. Search elements can be specified in random or sequential order. The matching documents are not ranked unless the modifier MANY (described below) is used.

The relational operators differ from the other operators in that they only apply to information in structured parts or fields of documents. (AUTHOR and DATE are possible fields.) When applying these operators, users must specify the field to be searched (called a FILTER in Verity's terminology), along with the name of the operator and the terms to be searched.

The relational operators are EQUALS, GREATER THAN, GREATER THAN OR EQUAL TO, LESS THAN, LESS THAN OR EQUAL TO, CONTAINS, ENDS, MATCHES, STARTS, SUBSTRING, and THROUGH.

The EQUALS operator searches a specific field and finds those documents whose values are identical to the search string.

The GREATER THAN operator searches a specific field and finds those documents whose values are greater than the search string.

The GREATER THAN OR EQUAL TO operator searches a specific field and finds those documents whose values are greater than or equal to the search string.

The LESS THAN operator searches a specific field and finds those documents whose values are less than the search string.

The LESS THAN OR EQUAL TO operator searches a specific field and finds those documents whose values are less than or equal to the search string.

The CONTAINS operator searches a specific field and finds those documents whose values

contain words or phrases in the query. Although the words in matching documents do not have to be contiguous, a user can set a parameter to CONTAINS that restricts matches to contiguous words. The CONTAINS operator does not recognize non-alphanumeric characters.

The ENDS operator searches a specific field and finds those documents whose values have the search string as an ending. For instance, a search of an author field using ENDS and the string "es" would return documents with the authors "Bates" and "Gates."

The MATCHES operator searches a specific field and finds those documents whose values match the search string. Only those values that match the search string are found. To loosen the rigidity of this operator, a question mark (?) or an asterisk (*) can be used to represent individual variable characters or multiple variable characters, respectively. For instance, MATCHES applied to "time" for a source field would return a document that contains precisely the string "time" and not "times" or "NY times." MATCHES applied to "time?" would find "time" and "times" but not "NY times?". The search "*time*" using MATCHES would find all three sources.

The STARTS operator behaves like ENDS but searches a specific field and finds those documents whose values have the search string as a prefix.

The SUBSTRING operator searches a specific field and finds those documents whose values contain a string matching the search term. The substring can be at the beginning, within, or at the end of a field value.

The THROUGH operator searches a specific field and returns documents containing values within the range of values in the search. Documents containing a date field, for example, can be searched for all the documents within a certain period of time.

In addition to the scoring process described above, some of these have impact on the returned list as follows:

If a child uses a Boolean operator (ANY or ALL), a proximity operator (PHRASE, SENTENCE, or PARAGRAPH), or a relational operator, the child receives a score of 1.00 if the topic is present, and 0.00 if not.

Modifiers can be associated with operators to alter their behavior. There are three modifiers, namely, CASE, MANY, and NOT.

The CASE modifier is used with WORD and WILDCARD operators for case sensitive searches. By default, searches are case insensitive.

The MANY modifier, which is used with several operators, counts the density of words, stems, or phrases in a document proportional to a document's length. It also produces a score for determining a document's relevance. The MANY modifier cannot be used with AND, OR, ACCRUE, or relational operators.

The NOT modifier excludes a specified word or phrase from a search.

Together the various operators can be used alone or in a number of combinations to support keyword, topic-based, or hybrid searches.

Operator Precedence

Another aspect of the operators is that they apply in a particular order, i.e., there is a precedence relationship that the operator used by a parent topic imposes on its offspring. Because they involve word searching in documents, WORD, STEM, SOUNDEX, WILDCARD, THESAURUS, and the relational operators take lowest precedence. The next level in the precedence hierarchy are, in ascending order, PHRASE, SENTENCE, PARAGRAPH, and ALL. The relationship here is based on the notion that a phrase is smaller than a sentence, paragraphs contain sentences, and documents (reflected by the ALL operators) consist of paragraphs.

The highest precedence are the concept operators AND, OR, and ACCRUE. The precedence of ANY depends on the parent topic.

Size of Indexing and Searchable Collections

Verity permits searching of 127 collections in a session where each collection can contain 8 million documents.

Indexing size is typically 30-50% of the original collection.

Differentiating Features

While Verity supports many of the same keyword capabilities as other IR vendors, the primary distinguishing functionality built into Verity's technology is the notion of a topic. Being structured groupings of meaningfully related information, topics encode knowledge or expertise of a domain. Once built, these structures can be used by searchers for doing topical searches without the explicit reliance on key word queries.

Virginia Systems Software Services, Inc.

Headquarters

Virginia Systems Software Services, Inc.
5509 West Bay Court
Midlothian, VA 23112

Tel: 804.739.3200
Fax: 804.739.8376
email: vasys@applelink.apple.com
WWW: Not available.

Corporate Fact Sheet

Business

Virginia Systems develops and markets information search and retrieval technology. It also bundles its information retrieval technology with OCR capabilities.

Markets

Virginia Systems markets its products primarily to lawyers, researchers, government agencies, educators, and publishers for use on PCs and Macintosh computers.

Products

Core Technology

Sonar Professional is a search and retrieval product for researching and annotating document collections. Virginia Systems markets another product called Sonar with fewer features than Sonar Professional.

Related Products

Sonar Image and Sonar Image Personal Edition offers scanning/OCR technologies with different levels of retrieval capabilities.

Sonar Bookends is an index generator that supports over 20 word processing and desktop publishing programs.

Sonar TOC for Quark XPress generates a table of contents for Quark XPress documents. Sonar TOC for Quark XPress requires Sonar Bookends.

Platforms

Sonar Professional runs on PCs under Windows, and Macintoshes, and Sun Workstations under OpenLook.

Benchmarks

Sonar Professional can search over 10,000 pages per second.

Technical Fact Sheet

Indexing

Inverted files are used as an index for searching. User have the option of using stop words as part of the index. The stop word list that comes with Sonar can be edited. Indexes can also include user-specified list of words or phrases.

Indexes and a table of contents can be generated automatically for either a single document or a collection of documents. The indexes can be hierarchical showing related topics. Page numbers and file names can be included in an index.

Search Features

Sonar Professional supports:

Boolean searches—Boolean operators are AND, OR, and NOT.

Proximity searches—Proximity is measured in terms of words.

Wild card searches—Words can be truncated for searching with wildcards.

Field/block searches—Searches can be restricted to specific parts of documents, e.g., titles.

Relevance ranking—Documents are ranked by number of hits.

Search summaries—Term highlighting at a page level is an option. Alternatively, users can view titles of returned documents.

Query expansion—A thesaurus is included and is used as a basis for generating alternative query terms.

Phonetic searches—Similar sounding words can be used for searching. This option relies on an

editable built-in dictionary of phonetically-similar words.

Search suspension—Users can suspend a search and begin another search. The suspended search can be resumed at the point of suspension.

Automatic hypertext linking of relationship—A user can view different documents mentioning the same subject through a special relationship mode in Sonar. In particular, relationships between people, places, and things is displayed through a hypertext link.

Associated word searches—Words found near a search term can be displayed by frequency or use, or in alphabetical order.

Commenting facility—Users can annotate hits (pages) with comments which will be displayed when the page is on the screen.

Conclusion

Although this survey is not competitive in that direct comparisons between individual vendors are avoided, a number of commonalities and differences are apparent, some of which might have more to do with marketing than technology. This section discusses some of the key marketing perspectives and or technology capabilities which are shared by some vendors and which distinguish others.

There are a number of search techniques that are fairly standard and well understood including Boolean searches, truncation operations, wildcard searches, term highlighting, database field searches, proximity searches, and relevance ranking. IR systems typically incorporate some number of the common search functions together with a few other capabilities (e.g., saved searches and wordwheels) that allows for easier and more effective searching by users. However, there are other types of functionality, namely, concept-based searching, fuzzy queries, query expansion, term weighting, and the ability to index and search large document collections, that are either fairly recent in inception, or have been available but only now play a central role in retrieval. Each of these expanded search functions will be discussed in the context of the IR vendors contained in this survey. There will also be a discussion of indexing methods, and how the notion of precision and recall is treated by IR vendors. Some general comments and observations will complete this section.

Concept-based Searching

The phrases "content-based retrieval" or "concept-based retrieval" are used in marketing literature as well as technical descriptions by many of the vendors surveyed. About half of the vendors offering IR systems on UNIX-based machines support content-based retrieval in some form. The notion of content-based or concept-based retrieval can be informally described as a method of searching for information that is based on meaning. The implication is that the technique uses some notion of semantic abstraction that goes beyond searching for terms explicitly present in the text, i.e., the information cannot be found by a keyword search. This loose definition fits a number of searching methods and satisfies all the surveyed vendors's notions of concept searching. However, although there are some aspects of concept searching that are shared by some vendors, in general, the way in which concept-based retrieval is implemented varies across vendors.

Concept-based retrieval presupposes some sort of semantic representation that reflects relationships between terms. These terms and their representations can be general or domain specific, manually created or machine generated, and products differ in terms of the types of representations and facilities for constructing the relationships. Verity, for instance, identifies concept-based searching with queries using topic trees that are manually constructed for a particular domain, whereas HNC uses context vector associations automatically derived from text as a basis for conceptual searching. IDI's notion of concept searching is founded on term expansion using an application-specific thesaurus that includes the notion of concept hierarchies, where the levels represent the degree of specificity of terms. ConQuest's major emphasis is on its dictionary, which contains word meanings and semantic relationships.

Although it can be extended for different domains, the semantic network that is delivered with each of ConQuest's products is general purpose. (Verity topics and IDI's concept hierarchies must be developed by the customer for each application.)

Fuzzy Searching

Fuzzy searching is another term that is used extensively in marketing and technology specifications of IR vendors. Moreover, the use of the technique is not restricted to the high-end systems; many PC- and Macintosh-based systems can perform fuzzy searching.

Fuzzy searching can be loosely described as a search technique that allows for term matches that are not exact. In practice, fuzzy searching has a variety of behaviors. For instance, in Concordance by Dataflight, fuzzy queries allow phonetic variants (e.g., "jim" and "gym") and orthographic variants (e.g., "US" and "U.S."). Excalibur supports letter substitution, transposition, and a form of stemming, while Odyssey's ISYS only supports letter substitution. Even more of a contrast is Fulcrum, which permits fuzzy Boolean searches, a relaxation technique primarily used for the Boolean operator AND, and having nothing to do with phonetic or orthographic variants.

Scalability

Most vendors claim that their technology is capable of indexing and searching large document collections, that is, their technology scales. In fact, many vendors identify the ability to scale as one distinguishing feature of their technology. Since the amount of on-line information is growing rapidly, scalability is certainly important and will likely be a necessary condition for an IR vendor to survive (even in the PC market). Here are some examples of sizes of document collections that can be indexed and searched.

CLARIT from CLARITECH can index up to 8GB of source text, while BASISplus (IDI) can handle a collection as large as 126GB. Moreover, since users can search up to 32 collections in a single session in BASISplus users, in principle, can search 4 Terabytes of data. ConQuest has been indexing and searching collections up to 100GB. Recently, ConQuest has been testing its system with 500GB collections. Open Text claims that there are no inherent size limitations on document collections for indexing with their system. Open Text has indexed 50GB collections. TextWare, a company with a personal computer product, can support 2GB of text in a single database.

Speed also plays a factor in scalability. Large collections must be indexed in reasonable periods of time. CLARIT can index 80-100 MB per hour on a DEC Alpha 3000/400 (133Mhz and 128MB RAM). On mid-range UNIX servers, BASISplus has an indexing throughput of up to 120MB per hour, while TextWare can index up to 32MB per hour on a 80-386 PC.

Even with the divergence in the size of collections that can be indexed and searched and the divergence in indexing time, it is clear that today's technology is capable of handling tens of gigabytes of text with a processing speed on the order of tens of megabytes per hour.

Query Expansion (Stemming and Thesaurus-based)

Query expansion refers to the process whereby a query is augmented by terms which are in some way related to the terms in the original query. Two methods that are commonly used are word-based (morphological) and thesaurus-based. In a system which supports stemming or morphology, inflected terms can be generated and used for uninflected terms in a query. For instance, a query with the simple term "tree" (an uninflected form) would be broadened so as to include a search for the plural form "trees" (the inflected form). A similar process can generate singular forms from plurals as well as third person singular, past tense, and participial forms of verbs. Such a system takes the burden off the searcher in remembering to use inflected forms or in formulating more complex forms containing inflected and uninflected forms.

Most systems do not use morphological processing or stemming, but give the user an option of using a wildcard. Proper use of wildcards can have approximately the same effect as stemming. However, wildcard searches place a demand on a searcher to introduce a wildcard in a term in a way that will not be too restrictive (documents will be missed) or too unrestrictive (irrelevant documents will be returned).

Another form of term expansion exploits a thesaurus. This method differs significantly from stemming and other expansion processes in that expansion is accomplished via semantic relationships rather than word structure operations. The process of semantic expansion involves a collection of words or phrases that are linked through a set of relationships including synonymy ("happy" and "glad"), antonymy ("happy" and "sad"), and isa ("birch" is a "tree"). Thus, a search for "happy" would also yield documents with the term "glad." Users can invoke any of the query expansion relationships (e.g., synonymy and antonymy) while doing a search in order to broaden searches as appropriate.

Thesaurus construction is a labor-intensive task and it is not surprising that few IR vendors offer a large thesaurus with their products. However, given the potential utility of thesaurus-based term expansion, vendors often support this type of expansion with user-supplied thesauri with the proper format (usually, ANSI Z39.19-1980 or ANSI/NISO Z39.19-1993). Cuadra is one such vendor that does not supply a thesaurus but incorporates query expansion functionality in its products.

ConQuest is one vendor that offers a large set of semantic relationships with its products for supporting term expansion. Currently, ConQuest network has over 400,000 concepts (terms) and 1.6 million word relationships.

Dataware, which has licensed and adapted technology from Inso, also includes thesaurus support as well as a semantic library API.

It is worthwhile to note that Inso's primary focus is to develop multi-lingual tools for supporting morphological processing and thesaurus functionality using a variety of techniques.

Term Weighting

In a query, usually each term is equal in value to any other in that one term match does not impact the scoring more than another during a search. However, certain systems allow users to

adjust weights, that is, assign preferences to particular terms in a query. By doing so, the user will influence the relevance ranking of hits, and documents will be prioritized differently than the case where all terms were of the same weight.

CLARIT, Excalibur, Fulcrum, and Verity all support term weighting. In each case, term weighting is performed manually.

Precision and Recall

The notion of precision and recall is the standard for evaluating IR systems. Academic papers reporting on novel IR techniques will use precision/recall as a measure for determining the effectiveness of the proposed techniques as compared with existing systems. In contrast to the research community, IR vendors do not provide precision/recall numbers for their systems. In fact, the terms precision and recall are occasionally used in the marketing literature of vendors, but no numbers are provided.

There are two exceptions, namely, CLARITECH and HNC. Both of these companies have participated in the ARPA-sponsored Text Retrieval Evaluation Conference (TREC), and will supply technical reports of their participation and results.

Indexing Methods

Each vendor surveyed employs some type of indexing process. Three of the vendors, namely, CLARITECH, Excalibur, and HNC use a vector model, Open Text uses suffix arrays, whereas the rest use some form of an inverted file for indexing. CLARITECH has a natural language approach to indexing and retrieval, while technology from Excalibur and HNC are purely string-based. The underlying work for Open Text evolved from work at the University of Waterloo and is unique in its general approach for commercial systems. It seems that the traditional inverted file method of indexing dominates the commercial market.

General Discussion

With 23 vendors surveyed, several observations stood out. The first obvious observation is that many vendors are either providing WWW search capabilities (Cuadra, Fulcrum, IDI, PLS, Thunderstone, and Verity) or are planning to do so (e.g., CLARITECH and ConQuest). Moreover, it is also clearly evident that a large percentage of vendors are supporting compound documents, i.e., documents that contain both text and non-text material such as images, sound, and other binary data. Vendors are also offering systems with seamless coupling of relational databases and textual databases, thereby providing users with access to data from textual and non-textual sources.

While vendors make marketing claims (along with case studies from satisfied customers) about the ease of use of their respective products, the focus is more on interface issues rather than ease of use of functionality inherent in search operations. For instance, Excalibur offers a File room metaphor and TextWare uses Cardfiles in order to help users organize, manage, and search their data. However, in these and other cases, the invocation of a particular operator is roughly identical with similar functionality to other systems and still leaves the burden on users to know when to use a particular operator and how to use it. Whereas some users prefer to control every aspect of each search, a system that eases the burden on the searcher and automatically adjusts

its searching in some way, provides an attractive alternative for many non-expert users. For example, "Fuzzy Boolean" searching, in which an AND of three terms is automatically relaxed to a 2-out-of-3 requirement if there are no matches for all three terms, saves the user a lot of query reformulation.

Since the 1960s, one type of technology that has been subject to a number of research projects regarding IR is natural language processing (NLP), where text is viewed as having linguistic structure rather than being a collection of strings. Morphological analysis, i.e., determining word structure (roots and affixes), and syntactic analysis, i.e., determining the hierarchical structure of language (noun and verb phrases and subjects and objects), are two components of NLP techniques that lend themselves to IR-related tasks. Of the companies surveyed, CLARITECH and Dataware offer natural language techniques or modules for use in extending traditional IR methods. One reason that NLP techniques are not included with IR systems is that specialized expertise is required to build and integrate NLP software. To help alleviate the resource costs of developing NLP software, Inso has developed a product suite of NLP technology aimed at the OEM market. Dataware and Fulcrum, both of which have previously offered a fairly standard set of searching capabilities, have been early buyers of Inso's technology in order to provide language-oriented approaches to retrieval.

Acknowledgements

I would like to thank Bill Woods who suggested conducting the survey and for reading the complete report in fine detail. His suggestions were invaluable. Gary Adams, Jacek Ambroziak, Patrick Martin, Philip Resnik, and Mark Torrance deserve thanks for commenting on the survey and answering questions on the technologies surveyed. Finally, the representatives from the various information companies also deserve thanks for taking the time to talk with me or return email about their companies and products.